



INDIAN AGRICULTURAL
RESEARCH INSTITUTE, NEW DELHI.

I.A.R.I.6.

MGIPC-SI-6 AR/54-7-7-54-10,000.

ANNALS
NEW YORK ACADEMY
OF SCIENCES

VOLUME XL



NEW YORK
PUBLISHED BY THE ACADEMY
1940

Editor
ERICH MAREN SCHLAIKJER

Assistant Editors
RICHARD P. HALL
(Pages 1-36, 133-266)
THEODORE SHEDLOVSKY
(Pages 37-132, 289-482)

CONTENTS OF VOLUME XL

	Page
Title Page.....	i
Contents.....	iii
Studies on the Nutrition of Colorless Euglenoid Flagellates. I. Utilization of Inorganic Nitrogen by <i>Astasia</i> in pure Cultures. By HENRY W. SCHOEN-BORN.....	1
Free Radicals as Intermediate Steps in the Oxidation of Organic Compounds. By L. FARKAS, MANUEL H. GORIN, L. MICHAELIS, OTTO H. MÜLLER, MAX-WELL SCHUBERT, and G. W. WHELAND.....	37
The Structure and Relationships of <i>Protoceratops</i> . By BARNUM BROWN and ERICH MAREN SCHLAIKJER.....	133
Recurrent Paleozoic Continental Facies in Pennsylvania. By BRADFORD WILLARD.....	267
Dielectrics. By WILLIAM O. BAKER, J. D. FERRY, RAYMOND M. FUOSS, PAUL M. GROSS, MARCUS E. HOBBS, JOHN G. KIRKWOOD, S. O. MORGAN, HANS MUELLER, J. L. ONCLEY, HERBERT A. POHL, J. SHACK, CHARLES P. SMYTH, and J. H. VAN VLECK.....	289

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

VOLUME XL, ART. 1. PAGES 1-36

MAY 15, 1940

STUDIES ON THE NUTRITION OF COLORLESS
EUGLENOID FLAGELLATES. I. UTILIZATION
OF INORGANIC NITROGEN BY ASTASIA
IN PURE CULTURES*

BY HENRY W. SCHOENBORN

Biological Laboratory, University College, New York University, New York

CONTENTS

	PAGE
INTRODUCTION	1
MATERIAL AND METHODS	5
EXPERIMENTAL RESULTS	8
Preliminary series with <i>Astasia</i>	8
Growth of <i>Astasia</i> in relation to light	10
Heteroautotrophic nutrition of <i>Astasia</i>	11
Growth of <i>Euglena gracilis</i> in media D, DH, DI, and DJ	13
Transfer of the heteroautotrophic strain of <i>Astasia</i> to other media	16
Growth of <i>Astasia</i> in relation to oxygen tension	18
Possible effect of iron and manganese on growth of <i>Astasia</i>	22
DISCUSSION	25
SUMMARY	32
LITERATURE CITED	33

INTRODUCTION

Investigations on food requirements of Protozoa have long engaged the attention of protozoologists. The development of the pure-culture technique and the establishment of steadily increasing numbers of Protozoa in pure cultures have, in recent years, added impetus to studies on the nutrition of these microorganisms. Such pure cultures,

* Awarded an A. Cressy Morrison Prize in Natural Science in 1939 by the New York Academy of Sciences. Accepted by the Graduate School of New York University in partial fulfillment of the requirements for the degree of Doctor of Philosophy. Publication made possible through a grant from the income of the George Herbert Sherwood Memorial Fund.

which are free from bacteria and all other organisms except the species under consideration, afford unique advantages in these investigations. The absence of other microorganisms eliminates such sources of error as the waste products of contaminants and the chemical and physical changes which might be produced in the medium by species other than the one in question. Pure cultures thus make possible, within experimental limits, the precise determination of food requirements and the exact measurement of effects produced on growth by different environmental factors.

The recent application of the pure-culture technique to studies on protozoan nutrition has led to a realization that the older classification—autotrophic, saprozoic and holozoic nutrition—is no longer adequate. Consequently, a more accurate designation of the various specialized types of nutrition has become necessary. These newer classifications attempt to define the simplest methods of nutrition possible for different types of Protozoa. The classification followed in the present paper is that given by Hall (1939b):

- I. Phototrophic nutrition.—energy of light is used in photosynthesis.
 1. Photoautotrophic nutrition.—inorganic nitrogen sources suffice for growth.
 2. Photomesotrophic nutrition.—amino acids are used as nitrogen sources.
 3. Photometatrophic nutrition.—exhibited by species capable of utilizing peptides or more complex nitrogen compounds as nitrogen sources.
- II. Chemoautotrophic nutrition.—inorganic media suffice for growth even in the absence of photosynthesis.
- III. Heterotrophic nutrition.—an organic carbon source is necessary. Observed mainly in colorless organisms, but some chlorophyll-bearing species may be considered facultative heterotrophs when grown in darkness.
 1. Heteroautotrophic nutrition.—inorganic nitrogen compounds are utilized.
 2. Heteromesotrophic nutrition.—one or more amino acids are used as nitrogen sources, and probably also as carbon sources; better growth is usually obtained with an additional carbon source.
 3. Heterometatrophic nutrition.—peptides or more complex nitrogen compounds are used as nitrogen sources.

According to this classification, a chlorophyll-bearing flagellate for which an amino-acid represents the simplest possible nitrogen source would be termed photomesotrophic; a similar organism, capable of using inorganic nitrogen when grown in light, would be termed photoautotrophic; and similarly, for the various other methods of nutrition. Photoautotrophs may be capable of photomesotrophic and photometatrophic nutrition, and photomesotrophs of photometatrophic nutrition; likewise, heteroautotrophs may be capable of carrying on heteromesotrophic and heterometatrophic nutrition, and heteromesotrophs capable

of heterometatrophic nutrition. However, the reverse is not true in either case.

One of the first groups whose nutritional requirements were investigated in pure cultures was the Euglenida. In this order photoautotrophic nutrition has been reported for several chlorophyll-bearing species. Although Zumstein (1899) and Ternetz (1912) concluded that inorganic media were unfavorable for *Euglena gracilis*, Pringsheim (1912) found growth of this species to be almost as heavy in certain inorganic media as in those containing organic nitrogen. More recently, Mainx (1928), Dusi (1933a), and Hall and Schoenborn (1939a) have described photoautotrophic nutrition in this species. Similar results have been obtained with *E. klebsii*, *E. stellata*, *E. anabaena* and *E. viridis* (Dusi 1933b; Hall 1938, 1939a). Hütner (1936) was unable to grow either *E. gracilis* or *E. anabaena* in certain inorganic solutions, but this may have been due to his use of unsatisfactory media, as shown by Hall and Schoenborn (1939a). Mainx (1928) reported that several additional Euglenidae were capable of utilizing inorganic nitrogen sources; however, Mainx did not use the serial-transfer technique and his conclusions must await confirmation. Certain of his findings have since been questioned. For example, Mainx considered *E. pisciformis* and *E. deses* photoautotrophic while later investigators, employing the method of serial transfers, have obtained negative evidence. It now appears that *E. deses* is photomesotrophic, and that *E. pisciformis* is photometatrophic (Dusi 1933b, Hall & Schoenborn 1938), or photomesotrophic if certain growth factors are furnished (Dusi 1939).

Attempts to grow green euglenoids in darkness have been none too successful, since *Euglena stellata*, *E. klebsii*, *E. anabaena*, *E. deses* and *E. pisciformis* have given negative results even in complex organic media (Dusi 1933b, Jahn 1935). Although heterometatrophic nutrition has been demonstrated in *E. gracilis*, inorganic nitrogen sources apparently have failed to support growth in darkness (Lwoff 1932, Dusi 1933a, Lwoff & Dusi 1934, 1937, 1938). *E. mesnili* also has been grown in darkness in peptone media (Lwoff & Dusi 1935a), but no attempt was made to culture this form on inorganic nitrogen sources.

In contrast to the observations on green Euglenida, little is known about the nitrogen requirements of colorless species. Pringsheim (1921) concluded that *Astasia ocellata* was incapable of utilizing amino acids or ammonium salts as nitrogen sources. Mainx (1928) later obtained growth of the same species in natural mixtures of amino acids (hydrolyzed proteins) but not on single amino acids or artificial

mixtures. Thus, the only Euglenida which have previously been found to utilize inorganic nitrogen salts are some of the chlorophyll-bearing species when grown in light. In darkness these forms, as well as the colorless Euglenida, have always appeared to require organic nitrogen sources.

In certain colorless Phytomonadida, however, the utilization of inorganic nitrogen salts has been reported by several investigators. Thus, Pringsheim (1921, 1934, 1935a) has described heteroautotrophic nutrition in *Polytoma uvella* and *Polytomella agilis*, and also in two chlorophyll-bearing species (*Chlorogonium euchlorum* and *C. elongatum*) grown in darkness. Later, however, this worker concluded (1935b) that continued growth of these species on inorganic nitrogen sources was possible only when "soil extract" was added to the medium. Furthermore, Lwoff and Lederer (1935) found that, in addition to any possible growth-factor effect, such a soil extract contained sufficient nitrogen to support growth of *Polytomella agilis*. More recently, Pringsheim (1937a, b) has replaced the soil extract with glucose-caramel, and has concluded that the latter is a necessary "Wuchsstoff" for growth of *Polytoma uvella*, *Polytomella (agilis) caeca* and *Chlorogonium euchlorum* (in darkness) on inorganic nitrogen sources. Lwoff and Dusi (1935b) have disagreed with Pringsheim's findings on *Chlorogonium euchlorum* and have concluded that this form is heteromesotrophic in darkness. Likewise, Loefer (1934) had previously reported that *C. euchlorum* and *C. elongatum* failed to grow in darkness on inorganic nitrogen sources. Hence, it is not certain that Pringsheim's strains of *Chlorogonium* actually carried on heteroautotrophic nutrition. On the other hand, Lwoff and Dusi (1938) have recently demonstrated heteroautotrophic nutrition of *Polytoma uvella* in media containing no growth factors, and this is again in disagreement with the latest findings of Pringsheim. Another phytomonad, identified as *P. uvella* by Lwoff (1932) but later designated as *P. obtusum* by Lwoff and Provasoli (1937), also appears to be heteroautotrophic (Lwoff 1932, Lwoff & Dusi 1938). For *Polytomella caeca*, Lwoff and Dusi (1938) found it necessary to add both the pyrimidine and thiazole components of thiamine (i.e., vitamin B₁) to the medium in order to maintain growth, while *P. caudatum* and *P. ocellatum* require only the thiazole constituent. *Hyalogonium klebsii*, another phytomonad, has been grown thus far only in peptone media and must therefore be considered heterometatrophic (Pringsheim 1937a).

The trophic nature of *Chilomonas paramecium*, the only species of Cryptomonadida studied up to the present time, is also in dispute.

This species has been reported as chemoautotrophic (Mast & Pace 1933), as heteroautotrophic (Mast & Pace 1933, Pringsheim 1934, 1935a), as capable of continuous growth in "heteroautotrophic" media only after soil extract is added (Pringsheim 1935b) or after certain growth factors are furnished (Lwoff & Dusi 1938), and as capable of heteromesotrophic nutrition (Hall & Loefer 1936).

It will be noted that there are definite differences of opinion concerning the simplest possible method of nutrition for many of the colorless flagellates thus far studied. The present investigation on *Astasia* was undertaken in order to determine the simplest media which will support growth of this colorless euglenoid flagellate. Although previous work indicated that both the colorless euglenoids and the green Euglenida grown in darkness are incapable of utilizing inorganic nitrogen sources, attempts were made to grow *Astasia* in such media. As a chlorophyll-bearing control species known to be capable of photoautotrophic nutrition, *Euglena gracilis* was also cultured in certain of the media in order to determine whether these solutions would support growth of green Euglenida. It was believed that such a study not only would determine the nitrogen requirements of *Astasia*, but would also throw some light on the factors responsible for the varied results obtained by different workers, even when studying the same organism.

The writer wishes to express his appreciation to Professor R. P. Hall who suggested the problem and who has given help and advice throughout the course of the investigation.

MATERIAL AND METHODS

The bacteria-free strain of *Astasia* sp. used in this investigation was originally isolated by Dr. T. L. Jahn and a specific name has not yet been proposed. The strain of *Euglena gracilis* was obtained in 1930 through the courtesy of Professor E. G. Pringsheim. Stock cultures of these species have since been maintained in our laboratory (Hall 1937a). The stock media employed in the present investigation were:

MEDIUM AA		MEDIUM AK	
Difco tryptone	. 10.0 gm.	Difco tryptone 7.5 gm.
KH ₂ PO ₄ ..	2 0	KH ₂ PO ₄ ..	1.0
Tap water	1 0 liter	NaC ₂ H ₅ O ₂ ·3H ₂ O	1.0
		Tap water	1.0 liter

The following experimental media were used:

MEDIUM EA	MEDIUM ECA
NH ₄ NO ₃	0.5 gm.
KH ₂ PO ₄	0.5
MgSO ₄ ·7H ₂ O.....	0.1
NaCl.....	0.1
FeCl ₃ ·6H ₂ O.....	0.0025
Dist. water.....	1.0 liter
MEDIUM EDA	MEDIUM EFA
(NH ₄) ₂ HPO ₄	1.0 gm.
MgSO ₄ ·7H ₂ O.....	0.2
KH ₂ PO ₄	0.2
KCl.....	0.2
FeCl ₃ ·6H ₂ O.....	0.0025
Dist. water.....	1.0 liter
MEDIUM D	MEDIUM DH
NH ₄ Cl.....	0.497 gm.
NaC ₂ H ₅ O ₂ ·3H ₂ O.....	1.148
MgSO ₄	0.048
K ₂ HPO ₄ ·3H ₂ O.....	0.209
Dist. water.....	1.0 liter

Medium EAB: Same as medium EA plus MnCl₂·4H₂O at 0.0001 gm. per liter.

Medium EF: Same as medium EFA plus MnCl₂·4H₂O at 0.0001 gm. per liter.

Medium DI: Same as medium EDA plus NaC₂H₅O₂·3H₂O at 1.0 gm. per liter.

Medium DJ: Same as medium EAB plus NaC₂H₅O₂·3H₂O at 1.0 gm. per liter.

All chemicals used were of analytical grade, and all were Mallinckrodt products with the exception of NaCl, (NH₄)₂HPO₄ and MgSO₄ (J. T. Baker), and NH₄Cl (Eimer and Amend).

The experimental procedure, that of successive transfers, has been used previously by Lwoff (1932), Dusi (1933a, b), Loefer (1934), Hall and Schoenborn (1939a) and others. A specific medium was prepared, adjusted to a suitable pH, measured into both 16 x 150 mm. tubes (9.0-9.5 cc.) and 125 cc. Erlenmeyer flasks (36-38 cc.), and then sterilized for 20 minutes at 115° C. Flasks were not used in the preliminary series; in the later series each flask always received exactly four times as much medium as did each tube. In making up a medium, the pH was adjusted by NaOH or HCl so that, after sterilization, it would be within the range 6.5-7.0. Good growth of *Astasia* sp. (Schoenborn 1936) and of *Euglena gracilis* (Jahn 1931, Dusi 1933a)

has been reported in this pH range. All pH readings were taken with a LaMotte Comparator.

Unless otherwise stated, the tubes and flasks of the first transfer in each series were inoculated from a culture in peptone medium. The tubes received either 0.5 or 1.0 cc. inocula, and the flasks four times as much as the tubes. Hence the initial concentration of flagellates was practically the same in both flasks and tubes. Three or four of the tubes thus inoculated were fixed for the initial count, one tube was used to determine the initial pH, and the remaining tubes and flasks were incubated at room temperature for periods specified below. At the end of the incubation period three or four tubes were fixed for the final count, one was used for determination of final pH, and a flask (or tube in the preliminary series) was used to inoculate the tubes and flasks of the second transfer. Similarly, a flask of the second transfer was used to inoculate the tubes and flasks of the third transfer, etc. The flagellate concentration in the tubes was determined by means of a Sedgwick-Rafter counting chamber and a Whipple micrometer (Hall, Johnson, & Loefer 1935). The counts thus obtained from tubes fixed at the beginning of the incubation period were averaged to give the initial count (x_0); those from tubes fixed at the end of the incubation period were averaged for the final count (x). Unless otherwise stated, the final population density of a flask culture was calculated from the initial count of the tubes inoculated from this flask. Growth is expressed as x/x_0 (ratio of final to initial concentration of flagellates per cubic centimeter) and, as used here, the term refers to the increase in number of flagellates (*i.e.*, growth of population).

In this method of serial transfers the peptone carried over from the stock medium was reduced, by serial dilution, to an insignificant amount in the later transfers. In the calculation of peptone concentrations (expressed as grams per cubic centimeter) in successive transfers, the peptone utilized by the flagellates was not considered; therefore the concentrations listed below are conservative estimates of the peptone present. Just as the peptone concentration is reduced by this method, so also will the concentration of flagellates be reduced unless a certain amount of growth occurs. The amount of growth necessary for a constant population level will depend, obviously, upon the technique of inoculation. In most cases, a 1.0 cc. inoculum was added to a tube containing 9 cc. of medium, with the result that the flagellate concentration was reduced to one-tenth in each transfer. Thus an average x/x_0 of 10 would be necessary in order to maintain

a given population level in successive transfers. If x/x_0 was lower than this, the concentration of flagellates would be reduced progressively until further transfers were impossible. This occurred in several series described below; the low growth rates of the flagellates made it necessary to discontinue these series after several transfers.

Pyrex tubes and flasks were used and all glassware was thoroughly washed and steamed, since chemical cleanliness is of utmost importance in studies on the nutritional requirements of Protozoa. Tubes, flasks and pipettes were plugged with bleached non-absorbent cotton (Johnson and Johnson) and then sterilized at 160° C. for three hours.

EXPERIMENTAL RESULTS

Preliminary Series with *Astasia*

The first six series with *Astasia* sp. were designed to obtain a good basic salt medium for later work. All these series were inoculated from stock cultures in medium AA. Only the initial pH values are given below, but the final pH never differed by more than 0.1 from the initial reading. The results obtained in these series are summarized in TABLE 1.

Series I (Medium EA). This series was carried through four transfers. The initial pH was 6.8 in the first and 6.7 in subsequent transfers. Initial counts in successive transfers were: 7,665; 1,188; 835 and 150; the periods of incubation were 14 days in the first two transfers and 21 days in the last two. Slow growth occurred in the first three transfers and none in the fourth (TABLE 1).

Series II (Medium EAB). Initial pH and periods of incubation were the same as for series I. Initial counts were: 7,665; 1,293; 765 and 268 in the successive transfers. The results (TABLE 1) were quite comparable to those obtained in series I.

Series III (Medium ECA). In this series, carried through four transfers, initial counts were: 6,485; 985; 655 and 68; initial pH: 6.9, 6.7, 6.7 and 6.8; and incubation periods: 15, 14, 21 and 21 days, respectively. Again slow growth occurred in the first three transfers and no growth in the fourth (TABLE 1).

Series IV (Medium EDA). The incubation periods were the same as for series III. Initial pH was 6.9, 6.7, 6.6 and 6.8; and the initial counts were: 6,485, 1,133; 763 and 110, respectively. The results (TABLE 1) show slightly better growth than in medium ECA of series III.

TABLE 1

SERIES I-VII. GROWTH IN TUBE CULTURES IN THE SUCCESSIVE TRANSFERS OF EACH SERIES IS EXPRESSED AS x/x_0 (RATIO OF FINAL TO INITIAL CONCENTRATION OF FLAGELLATES PER CUBIC CENTIMETER)

Series	x/x_0 in successive transfers					
	first	second	third	fourth	fifth	sixth
I	1.7	3.3	2.4	1.1	—	—
II	1.8	3.4	2.0	1.5	—	—
III	1.7	3.6	1.5	0.69	—	—
IV	1.7	4.4	2.1	1.4	—	—
V	1.5	2.8	1.6	—	—	—
VI	1.4	2.0	1.8	—	—	—
VIIa	2.0	1.7	—	—	—	—
VIIb	2.3	8.3	—	—	—	—
VIIc	2.7	11.9	3.8	5.9	1.4	1.4

Series V (Medium EFA). This series was carried through three transfers. The incubation periods were 14 days in the first two transfers and 21 days in the last. Initial pH was 6.9, 6.6 and 6.5 in the successive transfers, while the initial counts were: 6,815; 998 and 525. Very slow growth occurred in the three transfers (TABLE 1).

Series VI (Medium EF). In this series, carried for three transfers, initial pH and incubation periods were the same as in series V. Initial counts were 6,815 in the first transfer, 858 in the second, and 513 in the third. The growth rates (TABLE 1) were comparable to those in series V.

The growth observed in series I-VI may be attributed to the small amounts of peptone carried over in the inocula. Thus, the calculated peptone concentration was reduced from 1×10^{-3} grams per cubic centimeter in the first transfer to 1×10^{-6} in the fourth transfer, and growth of the flagellates was no longer evident when the peptone concentration was reduced to less than 1×10^{-5} ; hence, it was necessary to discontinue these series. So far as growth of *Astasia* is concerned, all of the media tested in these preliminary series were more or less comparable and none seemed to be toxic. Hence, the selection of media EDA and EAB for further studies (series XI, XII, XVI, XVII) was somewhat arbitrary.

Certain observations in these preliminary series suggested that light, a factor previously considered unimportant for colorless Euglenida, might exert a significant effect on growth of *Astasia*. In these series the tube cultures were incubated in test tube racks arranged on a tier of

shelves near a north window. However, all tubes were not exposed to equal intensities of light since, in order to conserve space, some tubes were placed behind others. It was noted that the tubes partially shielded by other tubes from the direct light of the window showed larger final populations and more nearly normal organisms than did the tubes nearer the window. In view of these results, it became necessary to determine more accurately the relation between light and growth of *Astasia*.

Growth of *Astasia* in Relation to Light

In testing the effects of light on growth of *Astasia*, tubes containing medium EDA were inoculated, divided into three groups, and incubated as follows: series VIIa, all tubes exposed to direct light from a north window; series VIIb, exposed to the same light filtered through a sheet of tracing paper; series VIIc, placed in light-proof containers adjacent to the other cultures. In these three series the initial pH was 6.7 in the first transfer and 6.9 in the second. In the first transfer, all three series were inoculated from the same stock culture, and the initial count in each case was 3,320. In the second transfer the initial counts were 417, 453 and 390 in series VIIa, VIIb and VIIc, respectively. Both transfers were incubated for 14 days. The results (TABLE 1) show that growth was better in filtered light than in direct light, and still better in darkness.

Since series I-VI had been carried out in light, which is not favorable for growth of *Astasia*, the series in darkness (VIIc) was continued for four additional transfers before the scarcity of flagellates made further work impossible. In these later transfers the initial pH was 6.7; the initial counts were 566, 153, 83 and 7; and the incubation periods 22, 40, 21 and 24 days, respectively. Results (TABLE 1) show continued slow growth during the third and fourth transfers, but the slight increase in numbers in the fifth and sixth transfers is of doubtful significance.

The results obtained in series VIIa, VIIb and VIIc indicate that growth of *Astasia* is inversely related to the intensity of light and that darkness is most favorable to growth. It is shown also that a purely inorganic medium, EDA, will not support growth of *Astasia* sp., even in darkness, when inoculated from a peptone stock culture.

With the selection of satisfactory inorganic media (EDA and EAB), the next step was to determine whether or not such solutions, with an added organic carbon source (e.g., sodium acetate) would support heteroautotrophic nutrition of *Astasia*. The results of series VII were

obtained during the progress of these series to be described, and the method of incubation was modified accordingly.

Heteroautotrophic Nutrition of *Astasia*

Several workers, mentioned above, have found that an organic carbon source, such as acetate, is necessary for growth of certain colorless phytomonad flagellates on inorganic nitrogen—in other words, these forms are heteroautotrophic and not chemoautotrophic. Accordingly, sodium acetate (1.0 gram per liter) was added to media EAB and EDA and these new media (designated as DJ and DI, respectively) were used in testing *Astasia* sp. for heteroautotrophic nutrition. In addition, medium D, a medium described by Mast and Pace (1933) for *Chilomonas paramecium*, was used, as well as one (medium DH) similar to that employed by Lwoff (1932) for *Polytoma obtusum*. With the exception noted below, these series were all inoculated from stock cultures in medium AK. Again, only the initial pH readings are given since the final pH values were never observed to differ by more than 0.2 pH unit from the initial readings, except in the first two transfers of those series started from peptone stock cultures. Results are summarized in FIGURE 1.

Series VIII (Medium D). In this series, carried through five transfers, initial pH was 7.1, 6.7, 6.7, 6.5 and 6.5, respectively; initial counts: 15,422; 1,933; 720; 153 and 10. Periods of incubation were 10, 14, 14, 14 and 15 days in the successive transfers. In view of the circumstantial evidence from series I–VI, that tubes shielded from the direct light of the window by other tubes showed more rapid growth and more nearly normal organisms than tubes not thus shielded, the cultures of this series were incubated with a sheet of tracing paper placed between the tubes and the window. The results (FIGURE 1) show that, under these conditions, medium D did not support continuous growth of *Astasia* sp.

Series IX (Medium DH). This medium differs from that used by Lwoff (1932) in that the FeCl_3 and CaCl_2 were added before autoclaving instead of after, and in that NaOH instead of K_2CO_3 was used to adjust the pH. Incubation periods for the first five transfers were the same as in series VIII, while the sixth and seventh transfers were incubated for 23 and 22 days, respectively. Initial pH was 6.9 in the first, 6.5 in the second, 6.4 in the third, and 6.5 in the remaining transfers; the initial counts were: 13,307; 965; 780; 216; 170 and 53 in the first six transfers, while in the seventh no organisms were seen when making the initial count. This series was also incubated in filtered

light. From the results (FIGURE 1) it may be concluded that continuous growth did not occur in this medium under the conditions specified.

Series X (Medium DH). The first transfer of this series was inoculated from the fifth transfer of series IX and was run concurrently with the sixth transfer of the latter. All transfers of series X were incubated in darkness, however, since results from series VIIa, VIIb and VIIc had now been obtained. The initial pH, periods of incubation, and initial counts were the same for the first and second trans-

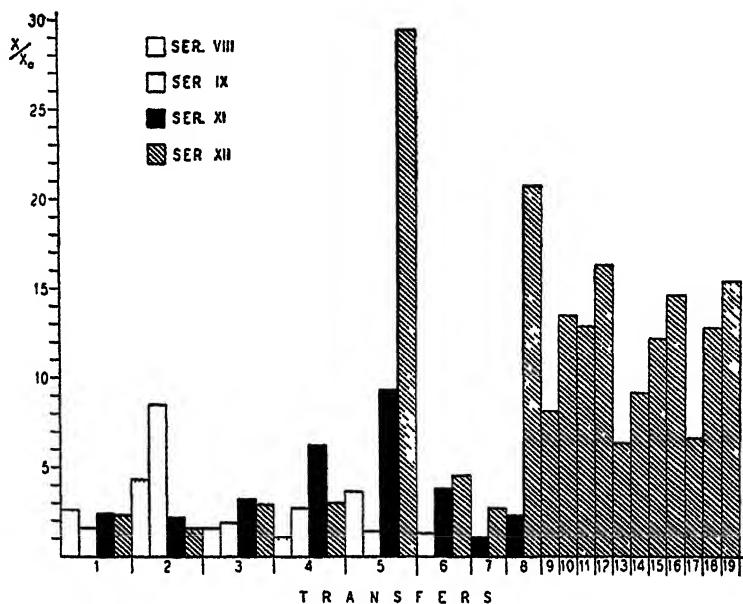


FIGURE 1 Series VIII, IX, XI and XII. Growth in tube cultures in successive transfers is expressed as x/x_0 (ratio of final to initial concentration of flagellates per cubic centimeter)

fers of series X as for the sixth and seventh transfers of series IX. The results were as follows: in the first transfer x/x_0 was 0.9, while in the second transfer no organisms were observed in making either the initial or final counts. Thus medium DH failed to support growth of *Astasia* sp. in darkness as well as in filtered light.

Series XI (Medium DJ). The initial pH was 6.9 in the first transfer and 6.5–6.7 in subsequent transfers. Periods of incubation for the first eight transfers were 4, 14, 14, 15, 23, 22, 21 and 21 days. The initial counts for these eight transfers were as follows: 23,720; 4,480; 903; 273; 123; 493; 200 and 33. After completion of the first four transfers, incubated in filtered light, the results of series VIIa, VIIb

and VIIc had been obtained; hence, all later transfers of series XI were incubated in darkness. The results from these first eight transfers (FIGURE 1) show slow growth in every transfer but the fifth, where slightly better growth occurred. In addition to the eight transfers described in FIGURE 1, this series was continued for two additional transfers in which no appreciable growth was observed. It then became impossible to carry this series any further.

Series XII (Medium DI). This series was carried through nineteen transfers, and the line is being continued at present as a stock strain. The first eight transfers were incubated concurrently with the first eight transfers of series XI and under the same lighting conditions. Later transfers (9-19 inclusive) were all incubated in darkness, each for 21 days except the eleventh which was allowed to grow 22 days. Initial pH was 6.9 in the first, 6.8 in the second, and 6.7 in all subsequent transfers. The initial counts for these nineteen transfers were: 23,720; 4,660; 630; 236; 96; 413; 152; 73; 157; 110; 190; 253; 386; 266; 353; 343; 290; 206 and 320, respectively. The calculated peptone concentration was decreased from 3.75×10^{-4} in the first transfer to 1.8×10^{-22} grams per cubic centimeter in the nineteenth. Results (FIGURE 1) show poor growth in the first four transfers, much heavier growth in the fifth, slow growth during the sixth and seventh, and then fair growth in all later transfers.

It was suspected at this time that the concentration of available iron in the medium might be one of the factors responsible for the fluctuations in growth observed, particularly in series XII. This question was investigated later and is discussed below in more detail.

Since it has now been possible to carry *Astasia* sp. through nineteen transfers in medium DI, covering a period of over a year, it is concluded that this organism is capable of heteroautotrophic nutrition under the conditions just described. However, the failure of this flagellate to grow in media D, DH and DJ was difficult to understand in view of the results obtained concurrently with *Euglena gracilis*.

Growth of *Euglena gracilis* in Media D, DH, DI, and DJ

In order to determine whether the media used for *Astasia* would support growth of one of the chlorophyll-bearing Euglenida, the same solutions were tested on *Euglena gracilis*, as described below. In every case, the cultures were incubated near a north window with the light filtered through tracing paper. With one exception as noted specifically below, these series were all started from a stock culture in medium

AK. No marked differences were observed between the initial and final pH readings.

Series XVIII (Medium DJ). In the first transfer, initial pH was 6.9; in the later transfers, 6.6–6.7. Periods of incubation for the successive transfers were as follows: 4, 21, 21, 23, 22, and 21 days in each of the last four. Initial counts were: 21,780; 10,186; 1,293; 96; 343; 113; 86; 130 and 116, respectively. Results (FIGURE 2) show slow

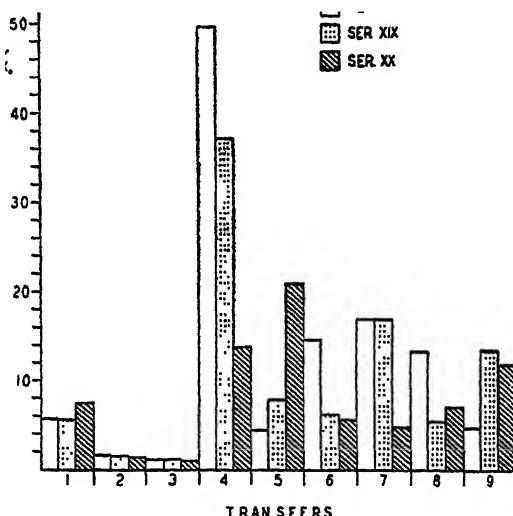


FIGURE 2. Series XVIII, XIX and XX. Growth of *Euglena gracilis* in tube cultures in successive transfers is expressed as x/x_0 (ratio of final to initial concentration of flagellates per cubic centimeter).

growth in the first transfer in which a considerable amount of peptone was undoubtedly still present, little or no growth in the second and third transfers in which the peptone concentration was reduced, and then a greatly increased growth rate in the fourth transfer. Subsequent transfers showed slow to moderate growth. The calculated peptone concentrations in transfers of this series (as well as for the next two series) are the same as for corresponding transfers of series XII. In the ninth and last transfer, this calculated value is quite low, 1.8×10^{-12} grams per cubic centimeter, and is probably not responsible for the growth observed. Thus, medium DJ apparently supports growth of *E. gracilis* in light.

The appearance of the organisms in various transfers of this series was interesting. In the first transfer the flagellates appeared normal;

in the second and third only a few normal green forms were seen, the majority being etiolated and otherwise abnormal in appearance; and then, at the end of the fourth transfer, all the flagellates observed were normal, green organisms. This normal type was predominant in all later transfers. Hence, it appears that selection, as previously described by Hall and Schoenborn (1939c) for *E. gracilis* when transferred to purely inorganic media, also occurred in this medium.

Series XIX (Medium DI). The initial pH and incubation periods were the same as in series XVIII. Initial counts for the nine successive transfers were: 21,780; 11,040; 1,413; 113; 210; 213; 136; 196 and 123, respectively. Results (FIGURE 2) are comparable to those obtained in series XVIII. Thus, this medium is capable of supporting continuous growth of *E. gracilis* and also exercises a selective action on the stock strain as did medium DJ.

Series XX (Medium D). In the first transfer the initial pH was 6.9 while in later transfers it was 6.5-6.7. The incubation period for each transfer was 21 days except for the first (10), second (14), fifth (23) and sixth (22). Initial counts for the transfers were: 35,588; 16,627; 2,013; 160; 233; 517; 363; 185 and 126. Results (FIGURE 2) are similar to those for the two preceding series. The increase in growth rate in the fourth transfer of this series was not as large as in series XVIII and XIX and some of the subsequent transfers also showed somewhat slower growth. However, selection apparently occurred, and the medium evidently will support growth of *E. gracilis*.

Series XXI (Medium DH). This series was not inoculated from a stock culture as were series XVIII to XX, but was started instead from the seventh transfer of series XX in which the calculated peptone concentration had already been reduced to 1.8×10^{-10} . In the five transfers of this series, initial pH was 6.5-6.7; all were incubated for 21 days except the fourth (22 days); and the initial counts were 185, 63, 266, 143 and 163. The x/x_0 values obtained were 4.2 in the first transfer, 36.7 in the second, 5.1 in the third, 19.2 in the fourth and 26.2 in the fifth. This series was discontinued after the fifth transfer in which there was a calculated peptone concentration of 1.8×10^{-15} grams per cubic centimeter. Hence, this medium also appears to be satisfactory for growth of *E. gracilis*.

With the completion of series XVIII-XXI, it was obvious that all four media would support growth of *Euglena gracilis*, in spite of the fact that medium DI alone had proven satisfactory for *Astasia*. The results obtained with *Euglena* suggested the possibility that the hetero-autotrophic strain of *Astasia*, which had been established in medium

DI, might be transferred successfully to the media which had failed to support growth of *Astasia* when inoculated directly from peptone stock cultures.

Transfer of the Heteroautotrophic Strain of *Astasia* to Other Media

Of the four "heteroautotrophic" media inoculated from peptone stock cultures, three (D, DH and DJ) had failed to support sufficient growth of *Astasia* to permit serial transfers. These three media were used in the following series to determine whether they would support growth of the heteroautotrophic strain of *Astasia* which had been established in medium DI. Each new series was started from a fairly late transfer of series XII. The results are shown in FIGURES 3 and 4. There were again no marked differences between the final and initial pH readings.

Series XIII (Medium D). The first transfer of this series was inoculated from the eighth transfer of series XII in which the calculated peptone concentration had already been reduced to 1.8×10^{-11} grams per cubic centimeter. The ten transfers of series XIII were incubated in darkness. Their initial pH was 6.5-6.7; the incubation period, 22 days in the third transfer and 21 in all others. Initial counts were 157, 140, 100, 133, 56, 213, 366, 222, 216 and 136, respectively. After the tenth transfer, in which the calculated peptone concentration had been reduced to 1.8×10^{-21} grams per cubic centimeter, series XIII was discontinued. Growth rates for the tube and flask cultures are described in FIGURE 3.

The growth rate in flask cultures was, on the average, distinctly higher than that in tube cultures. These results suggested a possible relationship between oxygen tension and growth of *Astasia*, since the relative surface area of the medium in flasks was more than ten times as great as that in tubes. Obviously, the oxygen tension of the medium in flask cultures would be appreciably higher than that in tube cultures. This question is discussed elsewhere in this paper.

Series XV. This series in medium DH, inoculated from the ninth transfer of series XII, was carried for three transfers before the scarcity of flagellates made further transfers impossible. The initial pH was 6.5 in all three transfers; periods of incubation were 21, 22 and 21 days; and initial counts were 110, 56 and 33, respectively. All transfers were incubated in darkness. Growth rates (x/x_0) were 2.9 in the first transfer, 2.8 in the second, and 1.8 in the third. This series confirms the results obtained in series IX and X, and demon-

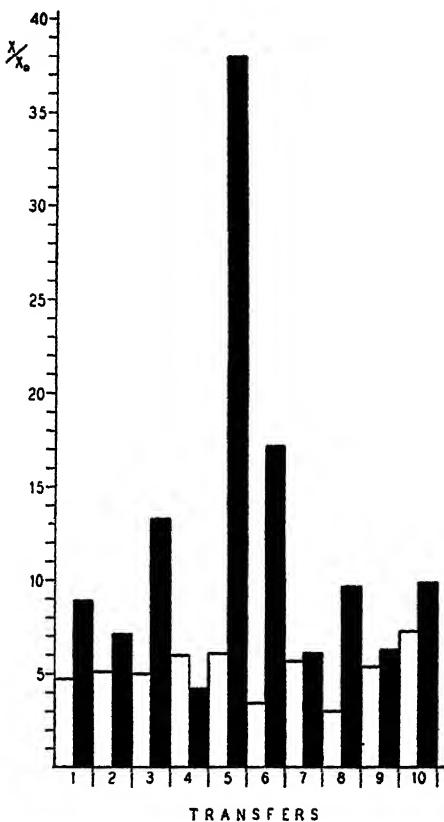


FIGURE 3. Series XIII. Comparison of growth in tube cultures (white blocks) and flask cultures (black blocks) in successive transfers. Growth is expressed as X/X_0 (ratio of final to initial concentration of flagellates per cubic centimeter).

strates that medium DH will not even support growth of the heteroautotrophic strain of *Astasia* sp. under the conditions specified.

Series XVI (Medium DJ). The ninth transfer of series XII, with a calculated peptone concentration of 1.8×10^{-12} , furnished the inocula for the first transfer in series XVI. In the ten transfers of this series the calculated peptone concentration was reduced to 1.8×10^{-22} grams per cubic centimeter. Each transfer was incubated in darkness for 21 days except the second (22 days); initial pH was 6.7 in all cases; and the initial counts were: 110, 103, 43, 36, 46, 76, 153, 293, 153 and 190, respectively. The results (FIGURE 4) show that poor to fair growth occurred in each transfer; thus, an autotrophic strain of *Astasia* is capable of continued growth in medium DJ.

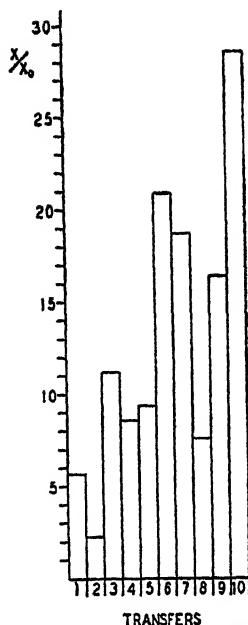


FIGURE 4 Series XVI Growth in tube cultures in successive transfers is expressed as X_0 (ratio of final to initial concentration of flagellates per cubic centimeter)

Of the three media which failed to support growth when inoculated from peptone stock cultures, only one (DH) has proven unsatisfactory when inoculated with the heteroautotrophic strain of *Astasia*. The other two media, D and DJ, appear to support continuous growth of this autotrophic strain. These results raise the question whether or not some process of adaptation is involved in the establishment of heteroautotrophic strains of *Astasia*. This problem is discussed in a later section of the paper.

It will be recalled that results with medium D indicated a possible relationship between the oxygen tension of the medium and the growth rate of *Astasia*. This observation led to further considerations of this factor.

Growth of *Astasia* in Relation to Oxygen Tension

As pointed out in the description of series XIII, differences between the amounts of growth observed in flask and in tube cultures suggested that oxygen tension of the medium might exert an important influence on growth of *Astasia*. Such a relationship between growth and oxygen tension has been reported previously for several other Protozoa in organic media. Thus, Rottier (1936a, b) obtained better growth of

Polytoma obtusum and *Polytomella agilis* in flask than in tube cultures, and attributed these results to a higher oxygen tension in the flasks. Furthermore, Jahn (1936) found that, after the first few days, growth of the ciliate, *Glaucoma piriformis*, was more abundant in aerated than in non-aerated peptone cultures. On the other hand, in the cryptomonad flagellate, *Chilomonas paramecium*, the results were reversed and both a lag period and a lower initial growth rate were observed in the aerated cultures.

In order to determine whether growth of *Astasia* in organic media also is influenced by oxygen tension of the medium, two series, XXII and XXIII, were carried out in a peptone solution (AK). Tube and flask cultures were exposed to the same environmental conditions, and direct counts were made on each to determine the final population densities. Both series were inoculated from a stock culture in medium AK, which had been seeded previously from the twelfth transfer of series XII; and both series showed an initial pH of 6.9. Series XXII, with an initial count of 2,276, was incubated in darkness for four days, and series XXIII, with an initial count of 653, was similarly incubated for six days. The results (TABLE 2) indicate that growth in flask cul-

TABLE 2

SERIES XXII AND XXIII. COMPARISON OF FINAL POPULATION DENSITIES AND OF GROWTH RATES IN TUBE AND FLASK CULTURES OF *Astasia* sp. GROWTH IS EXPRESSED AS x/x_0 (RATIO OF FINAL TO INITIAL CONCENTRATION OF FLAGELLATES PER CUBIC CENTIMETER), AND THE FINAL DENSITIES OF POPULATION AS THE NUMBER OF ORGANISMS PER CUBIC CENTIMETER

Series	Density of population		x/x_0	
	Tubes	Flasks	Tubes	Flasks
XXII	76,933	148,020	33.8	65.0
XXIII	111,920	283,070	171.4	433.5

tures was far heavier than that in tube cultures. This is especially true in series XXIII which had a lower initial count and was incubated longer than series XXII. In a peptone medium, therefore, it may be concluded that growth of *Astasia* sp. is definitely related to oxygen tension.

Whether this same conclusion is applicable to growth in "heteroautotrophic" media was determined by comparing the growth rates of flask cultures with those obtained in tube cultures. Five different media were employed. The results are summarized in FIGURE 3 and TABLE 3. No consistent differences between the growth rates of tube

The results obtained in peptone medium (series XXII and XXIII) and in medium D (series XIII) indicate that growth of *Astasia* may be increased by a favorable oxygen tension, both in a peptone solution and in a medium containing sodium acetate and an inorganic nitrogen source. On the other hand, no consistent differences in the growth rates of tube and flask cultures were observed in media DI and DJ, both of which are similar to medium D.

A comparison of the constituents of these three "heteroautotrophic" media shows that both DI and DJ contain an added iron salt, while the iron in medium D is limited entirely to the traces present as impurities in our samples of sodium acetate and potassium phosphate. Furthermore, medium DJ contains added manganese, an element not present as a detectable trace in medium D. Hence, it seemed that the deficiency in iron or in manganese, or in both elements, might be a clue to the different behavior of *Astasia* in medium D with respect to oxygen tension. The possible significance of iron and manganese was tested further, as described below.

Possible Effect of Iron and Manganese on Growth of *Astasia*

The possibility that growth of *Astasia* is accelerated by an adequate concentration of iron was tested in series XIV by the addition of $\text{FeCl}_3 \cdot 6\text{H}_2\text{O}$ (0.0025 grams per liter) to medium D. This series, carried for three transfers, was inoculated from the seventh transfer of series XIII (in medium D); thus the first transfer of XIV is comparable to the eighth of XIII, the second to the ninth, etc. In these transfers, a suitable quantity of medium D was prepared, and ferric chloride was added to half of the solution for series XIV. The other half, without iron, was tubed and used for series XIII. Thus, except for the ferric chloride, the media used for series XIII and XIV were identical. Likewise, all other factors except the initial counts were the same in transfers 1-3 of series XIV and transfers 8-10 of series XIII. The initial counts for the three transfers of series XIV were 222, 260 and 260. In FIGURE 5 the results obtained in series XIV are compared with those of the corresponding transfers in series XIII. It is obvious that the addition of iron resulted in higher population densities and growth rates in both flask and tube cultures. Another interesting result is that flask and tube cultures of series XIV showed fairly comparable growth rates, whereas the flask cultures of series XIII showed higher x/x_0 values, as pointed out above. Hence, the addition of iron to medium D overcomes the advantage of the flask culture attributable to its higher oxygen tension.

These results demonstrate that medium D does not contain the most favorable concentration of iron for growth of *Astasia*, and indicate further that iron is probably an essential element for this flagellate. Such a need for iron is readily understandable, since various substances important in cellular oxidations are known to contain iron. Thus, as a constituent, iron is necessary for the synthesis of cytochrome and the respiratory enzyme (cytochrome oxidase), both of which are considered as functional in respiration of aerobes. Peroxi-

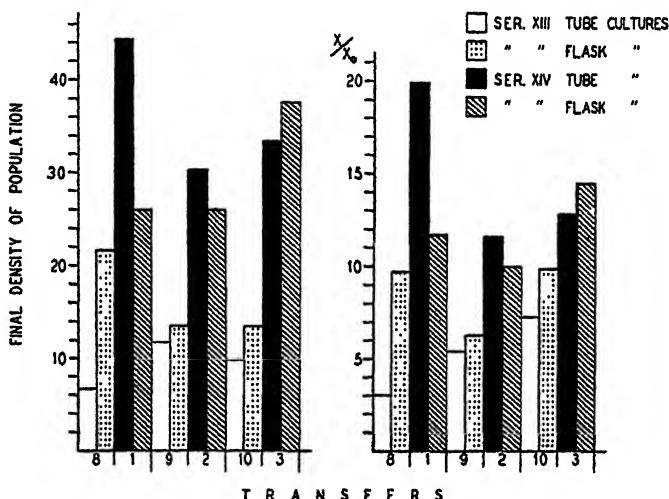


FIGURE 5. Comparison of final population densities (in hundreds per cubic centimeter) and growth rates (\times/\times_0 or ratio of final to initial concentration of flagellates per cubic centimeter) in tube and flask cultures of series XIII (transfers 8-10) and series XIV (transfers 1-3).

dase and catalase also contain iron. If any of these systems is operative in *Astasia*, iron would play an important role in cellular oxidation, a function suggested by some of the experimental results described above.

In the description of series XIII, it was suggested that some of the fluctuations in growth observed in this series might be correlated with an iron requirement of *Astasia*. In the preparation of medium DI for this series, a trace of iron was added by measuring 0.1 cc. of a 2.5% FeCl_3 stock solution into each liter of medium. This stock solution was kept in a dark cabinet and used for several transfers. A fresh solution was used in the fifth transfer of series XIII, and increased growth (FIGURE 1) was observed. This same solution was then used for transfers six and seven, and it is to be noted that the

growth rate dropped in the sixth transfer and still more in the seventh. Consequently, a fresh solution of ferric chloride was used in making media for the eighth and all later transfers. The eighth transfer also showed a decided increase in growth rate, and moderate growth has been observed in all later transfers of this series. The decreased growth rate when old iron solutions were employed may have different explanations. It has been maintained that, in old solutions of ferric chloride, some of the ferric ions would be precipitated as the hydroxide; and although autoclaving might induce such precipitation, nevertheless it seems likely that more ferric ions remain in the medium when a fresh rather than an old ferric chloride solution is used. Such an explanation may account not only for some of the observed differences in growth rates in series XII, but also for some of those irregularities observed in the early transfers of other series in which continued growth of *Astasia* was not obtained.

In all the series inoculated with the heteroautotrophic strain of *Astasia* (XIII, XV, XVI), a fresh ferric chloride solution was used in preparing media containing this substance. Thus, the observed differences in growth rates in the various transfers of these series were probably not due to the ferric ion, but to some other factor. Since temperature appears to be the only variable in these series, it may account for the results.

The possibility that manganese affects growth of *Astasia* was tested in series XVII by omitting the manganese chloride from medium DJ and comparing growth in this solution with that in medium DJ (series XVI). The three transfers of series XVII, the first of which was inoculated from the seventh transfer of series XVI, were comparable in all respects (except initial counts) to the last three transfers of the latter series, and the media for series XVI and XVII were identical except for the omission of manganese chloride in the latter. Initial counts were 293, 196 and 163. The observed final population densities of the three transfers of series XVII and the corresponding transfers of series XVI are: 2,080 (organisms per cubic centimeter) in the first transfer of series XVII as compared with 2,233 in series XVI (eighth transfer); in the second transfer, 1,640 as compared with 2,513; and in the third, 5,886 in comparison to 5,420. These differences are reflected in the growth rates in the various transfers (FIGURE 6): growth rates were similar in the first transfer; in the second, the medium with manganese supported better growth, and in the third transfer less growth.

It is obvious, therefore, that reduction of the manganese content of

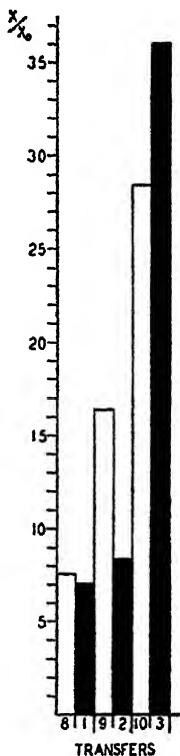


FIGURE 6. Comparison of growth in tube cultures of series XVI (transfers 8-10) and series XVII (transfers 1-3). Growth is expressed as χ/χ_0 (ratio of final to initial concentration of flagellates per cubic centimeter). Series XVI, white blocks; series XVII, black blocks.

medium DJ to a level approaching that of medium D exerts no significant effect on growth of *Astasia*. Hence, it seems that, if manganese is actually essential to this species, an adequate supply of this element is present as an undetected impurity in medium D, and also that the low concentration of manganese was not a factor limiting the growth of *Astasia* in this medium (series XIII) in tube cultures.

DISCUSSION

In addition to their intrinsic interest, investigations on the autotrophic nutrition of Protozoa have a definite bearing on certain more general phases of biology. Thus, the present observations on *Astasia* lead to the consideration of such problems as the fundamental chemical requirements of organisms, the question of acclimatization or adaptation to changing environmental conditions, and the factors in-

volved in the evolutionary origin of chlorophyll-bearing and colorless flagellates.

The extent of photoautotrophic and heteroautotrophic nutrition among the Protozoa is not yet known. At the present time, photoautotrophic nutrition has been demonstrated in five species of *Euglena* and also in five chlorophyll-bearing Phytomonadida. The remaining orders of the plant-like flagellates have not yet been investigated, although the discovery of photoautotrophic species in these other groups may well be expected. However, the presence of chlorophyll does not always insure photoautotrophic nutrition. Information now available indicates that certain chlorophyll-bearing flagellates are incapable of growth in inorganic media, as appears to be the case in *Euglena deses* and *E. pisciformis*. Hence, the loss of a primitive method of nutrition does not necessarily involve the disappearance of chlorophyll.

The known instances of heteroautotrophic nutrition are even less numerous than those of photoautotrophic nutrition. Thus, the heteroautotrophic nature of only one colorless phytomonad, *Polytoma obtusum*, remains undisputed at present, while the present paper describes the first instance of heteroautotrophic nutrition to be reported for the colorless Euglenida. The same method of nutrition has been reported also in *Polytoma uvella* (Lwoff & Dusi 1938), another phytomonad, and in *Chilomonas paramecium* (Mast & Pace 1933), a colorless cryptomonad flagellate. However, contradictory accounts in the literature make it necessary to withhold judgement until further evidence is forthcoming.

One of the striking features of the work on autotrophic nutrition is the frequent failure of investigators to obtain comparable results, even with the same species. Hence, the trophic nature of several colorless and chlorophyll-bearing flagellates is still in dispute. The occurrence of such differences in experimental results naturally suggests that the technical problems involved in such investigations are difficult ones. In the present study of *Astasia*, for example, the establishment of heteroautotrophic strains was complicated by problems involving inhibitory effects of light on growth of a colorless organism, by problems involving oxygen tension, by a deficiency of iron in one medium, and by an apparent process of adaptation or acclimatization of the flagellates to a simple medium. It seems probable that similar difficulties may have been encountered by other workers and may have been responsible in part for some of the conflicting results reported in the literature.

Another technical difficulty in these investigations involves the method of inoculation followed in the serial-transfer technique. As pointed out in the description of material and methods, the growth of a strain must compensate for the dilution of organisms which occurs in the inoculation of every transfer. Otherwise, the rate of dilution will exceed the rate of multiplication and the strain can be carried for only a few successive transfers. If the dilution at inoculation is 1 : 10, as in the present work on *Astasia*, the growth rate (x/x_0) in each transfer must involve, on the average, a ten-fold increase in number if the strain is to survive a series of successive transfers. On the other hand, certain investigators (e.g., Lwoff & Dusi) have used inocula of one or two drops. Obviously, a one-drop inoculum would be diluted approximately 200 times when added to 10 cc. of culture medium; a two-drop inoculum, about 100 times. In the first case the strain must maintain a growth rate approximating 200 if it is to be carried for a number of transfers; with two-drop inocula, x/x_0 must approach 100 under similar conditions. In the present investigation, the growth rate of heteroautotrophic strains of *Astasia* in tube cultures never exceeded 30 for any transfer, and the average for all transfers was much lower. Consequently, with one-drop inocula instead of the usual 1.0 cc., only a few transfers might have been possible, and negative rather than positive conclusions might have been reached concerning the heteroautotrophic nature of *Astasia* sp. It is possible, as suggested by the observations of Hall and Schoenborn (1939b) on *Euglena*, that a higher growth rate of *Astasia* might have been obtained with inocula smaller than 1.0 cc. Whether such a growth rate could approach x/x_0 values of 100 or 200 is problematical, however. At any rate, the results of Lwoff and Dusi (1938) must be reexamined on the basis of larger inocula, for the possibility remains that some or all of those species described as incapable of heteroautotrophic nutrition may actually grow very slowly in inorganic-nitrogen media, thus making serial transfers impossible when small inocula are used. This same criticism may also be applicable to the work of Pringsheim (1937a, b) since he states that a "Wuchsstoff" is necessary, *at least with small inocula and repeated transfers in series*. It seems possible, therefore, that future investigations may lead to revision of current views and that some of the other colorless flagellates may actually be comparable to *Astasia* in the ability to carry on heteroautotrophic nutrition.

The occurrence, in *Astasia*, of adaptation to simple media is suggested by the fact that it was possible to establish an autotrophic

strain in only one medium (DI) inoculated directly from a stock culture in peptone medium; yet two other media were found to support growth of the heteroautotrophic strain from medium DI, although they were unsatisfactory for the flagellates from peptone medium.

A similar adaptive process has been described in strains of *Euglena gracilis*, *E. anabaena* and *E. deses* when transferred to purely inorganic media (Hall & Schoenborn 1939c). In the present experiments with *E. gracilis*, in which the media contained an organic carbon source in addition to the inorganic constituents, this same type of adaptation has been observed. These cases of adaptation in *Euglena* are characterized, during the first few transfers in new media, by a marked decrease in growth rate and by the appearance of many etiolated and moribund flagellates. It may be assumed that in our stock strains of *Euglena* only a small percentage of the flagellates is capable of photoautotrophic nutrition or of becoming adapted to inorganic media; thus, when the strain is transferred into an inorganic medium it is only this small percentage which continues to live. These few flagellates remain green and healthy in appearance, and they continue to divide. However, their increase in number is overshadowed by the death of those flagellates incapable of growth in the new medium. Finally, only the green autotrophic flagellates remain, and the growth rate apparently increases.

Various other instances of adaptation, or acclimatization, have been reported for Protozoa as well as for bacteria. For example, some species may become acclimated to media of increased osmotic pressure and even to salinities equivalent to that of sea water (Loefer 1939). Other examples involve adaptation to chemicals (e.g., Jollos 1921), and to abnormally high temperatures (Dallinger 1907, cited by Calkins 1933).

Such adaptations are open to two explanations. In the first place, it may be assumed that the environmental change induces a real physiological change in the organism, and that this physiological modification persists as long as the new conditions are maintained. Or, the process of adaptation may involve selection, within a strain, of those organisms capable of survival under the new conditions, while the rest of the organisms die. The available evidence indicates that the second explanation may be applicable to *Euglena*. The apparent absence of morphological changes in *Astasia* makes it impossible to conclude definitely that selection is involved in the instance described, although it is obvious that selection might occur without any morphological indications. Likewise, in the ab-

sence of any adequate evidence, the writer is unwilling to assume that the establishment of a heteroautotrophic strain of *Astasia* involves the induction of a true change in metabolic nature. For the present, therefore, the mechanism involved in the apparent adaptation of *Astasia* to simple media must remain undetermined.

In addition to the general interest attached to such phenomena, the occurrence of adaptations may aggravate technical difficulties in the work on protozoan nutrition. If such processes are operative in other flagellates as well as in *Euglena* and *Astasia*, they may help to explain some of the existing controversies in regard to the trophic nature of the plant-like flagellates.

The establishment of photoautotrophic and heteroautotrophic strains of Protozoa makes possible for the first time a detailed consideration of the chemical requirements of these organisms. Such information is sadly lacking, and it is not yet possible to list with certainty the elements essential to metabolism and growth of any protozoon. The demonstration of photoautotrophic and heteroautotrophic nutrition thus represents the first basic step in the determination of such requirements.

The use of analyzed reagents has already made it possible to grow Protozoa in media whose composition is definitely known, within the limits of such analyses. On this basis, it is now possible to compile tentative lists of elements which should include all those essential to life of the plant-like flagellates. For example, medium DI, in which the original heteroautotrophic strain of *Astasia* was isolated, contains the following elements: C, H, O, N, P, K, Mg, S, Na and Cl in appreciable amounts, as well as traces of As, Cu, Fe, Mn, Ca, Pb, Zn and Ba. Two other media (D and DJ), which supported growth of the heteroautotrophic strain after it was once established, differ from medium DI in the following respects: As, Cu, Mn, Zn and Ba are not present as detectable traces in medium D and thus may not be necessary for *Astasia*; in medium DJ there is a trace of Al but none of Ba. Such lists probably include all, or at least nearly all of the elements essential for growth of such organisms as *Astasia*. While it is not impossible that a few other elements, present as impurities in undetected traces, may also be essential, this does not seem likely, since the concentration of any such elements in the media probably would not exceed 1×10^{-15} or 1×10^{-14} grams per cubic centimeter.

On the other hand, it is by no means certain that all of the elements included in these lists are essential to life of the plant-like flagellates. Some of them are obviously essential: for example, C, H, O and N,

as fundamental constituents of protoplasm. It may also be assumed that Ca, Cl, Fe, K, Mg, Na, P and S are probably essential, since all of these are more or less common in living cells. In addition, there is confirmatory evidence indicating that certain of these elements are essential. Magnesium, as a constituent of chlorophyll, is undoubtedly essential for the chlorophyll-bearing organisms, but this does not imply that the element is necessarily important in the metabolism of colorless flagellates. The essential nature of iron is suggested by some of the present observations on *Astasia*, and similar findings have been reported for *Polytoma obtusum* (Lwoff 1930). A need for calcium has been demonstrated in *Euglena stellata* (Dusi 1933b) and in the colorless phytomonad, *Hyalogonium klebsii* (Pringsheim 1937a). Manganese in low concentrations accelerates growth of *Euglena ambae* (Hall 1937b), but there is as yet no evidence that this element is essential to growth of colorless flagellates; observations on *Astasia* indicate that it is not essential, unless undetectable traces are adequate for growth of this species. Sulphur appears to be essential to growth of the cryptomonad, *Chilomonas paramecium* (Mast & Pace 1935). So far, however, there is no experimental evidence to suggest that Cl, K, Na or P are really essential to growth of the plant-like flagellates.

Other elements, which are present in traces in the media used for *Astasia*, may or may not be essential to growth. No experimental evidence bearing on the significance of these elements is yet available.

This brief discussion indicates the general lack of information concerning the qualitative chemical requirements of Protozoa. Here and there, fragmentary observations have indicated that certain elements are essential, and there are clues to the physiological significance of particular elements in a few instances. However, knowledge of the qualitative requirements of any single species is still incomplete. Furthermore, practically nothing is known about the quantitative needs of Protozoa. Thus, optimal, minimal and maximal concentrations are yet to be determined for practically every element found in the media which support the different types of autotrophic nutrition.

A more detailed body of knowledge concerning qualitative and quantitative chemical requirements is essential to further investigations in various fields of protozoology. For example, the work on the vitamin, or "growth-factor," requirements of Protozoa has been carried out with complete disregard for the underlying inorganic requirements of the species and in almost total ignorance of such requirements—all this in spite of significant observations hinting at complementary relationships between vitamins and inorganic elements

and suggesting the possible importance of certain elements in the synthesis of vitamins by the organism. In fact, it is not unreasonable to suspect that much of the work on protozoan growth-factors will have to be repeated after more nearly adequate knowledge of inorganic chemical requirements has been obtained.

The accumulation of precise information concerning inorganic requirements furnishes the only sound basis for the extension of studies on simple methods of nutrition in Protozoa. It is obvious that no flagellate will grow in a simple medium which does not satisfy, both qualitatively and quantitatively, the chemical requirements of the species. Until more is known about such requirements, there will be many failures to establish organisms in the simpler types of media and the results will be interpreted as indicating a need for proteins or perhaps for specific vitamins, whereas the underlying factor may be a simple deficiency in an inorganic element.

The demonstration of heteroautotrophic nutrition in flagellates was, in itself, surprising. But the very existence of heteroautotrophs suggests the interesting possibility that still simpler methods of nutrition may be possible in the colorless plant-like flagellates. In other words, it is not inconceivable that some of the colorless flagellates may actually be chemoautotrophic—that is, capable of growth in purely inorganic media and of obtaining energy by the oxidation of inorganic compounds alone. Mast and Pace (1933) have reported that such is true in *Chilomonas paramecium*; however, Pringsheim (1935a) was unable to confirm this observation, and several other investigators have been unable to grow this form even in media containing an organic carbon source (*e.g.*, acetate) in addition to the inorganic constituents. Thus, it is not certain that *C. paramecium* is actually chemoautotrophic. It is not illogical to speculate on the existence of this type of nutrition in colorless flagellates, however, since the nitrifying bacteria have been grown in inorganic media similar to those which support the chlorophyll-bearing photoautotrophic flagellates. Certain of these bacteria oxidize ammonium salts as the source of energy and thus do not require organic compounds or the energy of light. The heteroautotrophic flagellates utilize ammonium salts in media which, except for the presence of sodium acetate, are comparable to those supporting growth of the chemoautotrophic nitrifying bacteria. Hence, it seems possible that, with the development of satisfactory inorganic media, growth of such flagellates may even be obtained in the absence of an organic carbon source such as acetate.

The discovery of such a chemoautotroph among the flagellates

would be of interest not only as a contribution to general physiology but also in its bearing on the evolution of the flagellates. It is generally assumed that the chlorophyll-bearing plant-like flagellates are the most primitive of the Protozoa, and that further evolution has involved the loss of chlorophyll accompanied by both physiological and morphological specialization. The demonstration of heteroautotrophic nutrition in several flagellates has already shown that nutrition may be simple in the absence of chlorophyll. Furthermore, loss of the ability to use inorganic nitrogen compounds has apparently occurred in certain chlorophyll-bearing flagellates (e.g., *Euglena deses*, *E. pisciformis*). Hence, the presence of chlorophyll is no guarantee that nutrition is simple. The discovery of chemoautotrophic flagellates might even suggest that chlorophyll represents an acquisition in the later evolution of flagellates from primitive colorless ancestors. The desirability of searching for chemoautotrophs is obvious, but extreme caution must be exercised both in technique and in interpretation of experimental results. Successful results, if they are to be forthcoming at all, will depend upon more adequate information concerning the inorganic food requirements of Protozoa.

SUMMARY

It has been demonstrated that *Astasia* sp. is capable of heteroautotrophic nutrition; that is, this flagellate grows in media containing inorganic nitrogen and organic carbon sources. The establishment of heteroautotrophic strains has been complicated by several factors: (1) light has been shown to have an inhibitory effect on growth of this colorless species; (2) a relationship has been observed between growth and the oxygen tension of certain media; (3) the importance of iron to growth of *Astasia* has been demonstrated; and (4) it appears that a process of adaptation takes place when this species is transferred from peptone stock cultures to simple media. The control organism, *Euglena gracilis*, grew in the same solutions and confirmation was obtained that a process of adaptation, involving selection, occurs when the stock strain of *E. gracilis* is transferred to such media.

LITERATURE CITED

- Calkins, Gary Nathan, 1869—
- 1933. The Biology of the Protozoa. Philadelphia, Lea and Febiger 607 pp.
- Dusi, Hisatake
- 1933a. Recherches sur la nutrition de quelques Euglènes. I. *Euglena gracilis*. Ann. Inst. Pasteur 50: 550-597.
 - 1933b. Recherches sur la nutrition de quelques Euglènes. II. *Euglena stellata*, *klebsii*, *anabaena*, *deses*, et *pisciformis*. Ann. Inst. Pasteur 50: 840-890.
 - 1939. La pyrimidine et le thiazol, facteurs de croissance pour le flagellé à chlorophylle, *Euglena pisciformis*. Compt. Rend. Soc. Biol. 130: 419-422.
- Hall, Richard Pinkham, 1900—
- 1937a. "Growth of free-living Protozoa in pure cultures" in Culture Methods for Invertebrate Animals. Ithaca, Comstock 51-59.
 - 1937b. Effects of manganese on the growth of *Euglena anabaena*, *Astasia* sp. and *Colpidium campylum*. Arch. Protistenk. 90: 178-184.
 - 1938. Nitrogen requirements of *Euglena anabaena* var. *minor*. Arch. Protistenk. 91: 465-473.
 - 1939a. The trophic nature of *Euglena viridis*. Arch. Zool. expér. gén. 80: 61-67.
 - 1939b. The trophic nature of the plant-like flagellates. Quart. Rev. Biol. 14: 1-12.
-
- ; & Loefer, John Benjamin, 1908—
- 1936. On the supposed utilization of inorganic nitrogen by the colorless cryptomonad flagellate, *Chilomonas paramecium*. Protoplasma 26: 321-330.
-
- ; & Schoenborn, Henry William, 1912—
- 1938. Studies on the question of autotrophic nutrition in *Chlorogonium euchlorum*, *Euglena anabaena* and *Euglena deses*. Arch. Protistenk. 90: 259-271.
 - 1939a. The question of autotrophic nutrition in *Euglena gracilis*. Physiol. Zool. 12: 76-84.
 - 1939b. Fluctuations in growth-rate of *Euglena anabaena*, *E. gracilis*, and *E. viridis*, and their apparent relation to initial density of population. Physiol. Zool. 12: 201-208.
 - 1939c. Selective effects of culture media in bacteria-free strains of *Euglena*. Arch. Protistenk. 93: 72-80.
-
- ; Johnson, David Franklin, 1909— ; & Loefer, John Benjamin, 1908—
- 1935. A method for counting Protozoa in the measurement of growth under experimental conditions. Trans. Amer. Micr. Soc. 54: 298-300.
- Hutner, S. H.
- 1936. The nutritional requirements of two species of *Euglena*. Arch. Protistenk. 88: 93-106.
- Jahn, Theodore Louis, 1905—
- 1931. Studies on the physiology of the euglenoid flagellates. III. The effect of hydrogen ion concentration on the growth of *Euglena gracilis* Klebs. Biol. Bull. 61: 387-399.

1935. Studies on the physiology of the euglenoid flagellates. VI. The effects of temperature and of acetate on *Euglena gracilis* cultures in the dark. Arch. Protistenk. 86: 251-257.
1936. Effect of aeration and lack of CO₂ on growth of bacteria-free cultures of Protozoa. Proc. Soc. Exper. Biol. Med. 33: 494-498.
- Jollos, Victor**
1921. Experimentelle Protistenstudien. I. Untersuchungen über Variabilität und Vererbung bei Infusorien. Arch. Protistenk. 43: 1-222.
- Loefer, John Benjamin, 1908-**
1934. The trophic nature of *Chlorogonium* and *Chilomonas*. Biol. Bull. 66: 1-6.
1939. Acclimatization of fresh-water ciliates and flagellates to media of higher osmotic pressure. Physiol. Zool. 12: 161-172.
- Lwoff, André, 1902-**
1930. Le fer, élément indispensable au flagellé *Polytoma uvella* Ehr. Compt. Rend. Soc. Biol. 104: 664-666.
1932. Recherches biochimiques sur la nutrition des Protozoaires. Monographies de l'Institut Pasteur. Paris, Masson 158 pp.
- ; & Dusi, Hisatake
1934. L'oxytrophie et la nutrition des flagellés leucophytes. Ann. Inst. Pasteur 53: 641-653.
- 1935a. La suppression expérimentale des chloroplastes chez *Euglena mesnili*. Compt. Rend. Soc. Biol. 119: 1092-1095.
- 1935b. La nutrition azotée et carbonée de *Chlorogonium euchlorum* à l'obscurité; l'acide acétique envisagé comme produit de l'assimilation chlorophyllienne. Compt. Rend. Soc. Biol. 119: 1260-1263.
1937. Le thiazol, facteur de croissance pour *Polytoma ocellatum* (Chlamydomonadiné). Importance des constituants de l'aneurine pour les flagellés leucophytes. Compt. Rend. Acad. Sci. 205: 882-884.
1938. Culture de divers flagellés leucophytes en milieu synthétique. Compt. Rend. Soc. Biol. 127: 53-56.
- ; & Lederer, Edgar
1935. Remarques sur l'"extrait de terre" envisagé comme facteur de croissance pour les flagellés. Compt. Rend. Soc. Biol. 119: 971-973.
- ; & Provasoli, Luigi
1937. Caractères physiologiques du flagellé *Polytoma obtusum*. Compt. Rend. Soc. Biol. 126: 279-280.
- Mainz, Felix, 1900-**
1928. Beiträge zur Morphologie und Physiologie der Eugleninen. II. Untersuchungen über die Ernährungs- und Reizphysiologie. Arch. Protistenk. 60: 355-414.
- Mast, Samuel Ottmar, 1871- ; & Pace, Donald Metcalf, 1906-**
1933. Synthesis from inorganic compounds of starch, fats, proteins and protoplasm in the colorless animal, *Chilomonas paramecium*. Protoplasma 20: 326-358.

1935. Relation between sulphur in various chemical forms and the rate of growth in the colorless flagellate, *Chilomonas paramectum*. *Protoplasma* 23: 297-325.

Pringsheim, Ernst Georg, 1881-

1912. Kulturversuche mit chlorophyllführenden Mikroorganismen. II. Zur Physiologie der *Euglena gracilis*. *Beit. Biol. Pflanz.* 12: 1-47.
1921. Zur Physiologie saprophytischer Flagellaten (*Polytoma*, *Astasia*, und *Chilomonas*). *Beit. allg. Bot.* 2: 88-138.
1934. Über die pH-Grenzen einiger saprophytischer Flagellaten. *Naturwiss.* 22: 510.
- 1935a. Über Azetatflagellaten. *Naturwiss.* 23: 110-114.
- 1935b. Wuchsstoffe im Erdboden? *Naturwiss.* 23: 197.
- 1937a. Beiträge zur Physiologie saprophytischer Algen und Flagellaten. I. *Chlorogonium* und *Hyalogonium*. *Planta, Arch. wiss. Bot.* 26: 631-664.
- 1937b. Beiträge zur Physiologie saprophytischer Algen und Flagellaten. II. *Polytoma* und *Polytomella*. *Planta, Arch. wiss. Bot.* 26: 665-691.

Rottier, P.-B.

- 1936a. Recherches sur les courbes de croissance de *Polytoma uvella*. L'influence de l'oxygénation. *Compt. Rend. Soc. Biol.* 122: 65-68.
- 1936b. Recherches sur la croissance de *Polytoma uvella*. L'influence de la concentration des substances nutritives. *Compt. Rend. Soc. Biol.* 122: 776-780.

Schoenborn, Henry William, 1912-

1936. Growth of two species of *Astasia* in relation to pH of the medium. *Anat. Rec.* 67 suppl.: 121.

Ternetz, Charlotte

1912. Beiträge zur Morphologie und Physiologie der *Euglena gracilis* Klebs. *Jahrb. wiss. Bot.* 51: 435-514.

Zumstein, Hans

1899. Zur Morphologie und Physiologie der *Euglena gracilis* Klebs. *Jahrb. wiss. Bot.* 34: 149-196.

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

VOLUME XL, ART. 2. PAGES 37-132

JUNE 20, 1940

FREE RADICALS AS INTERMEDIATE STEPS IN THE OXIDATION OF ORGANIC COMPOUNDS

By

L. FARKAS, MANUEL H. GORIN, L. MICHAELIS, OTTO H. MÜLLER,
MAXWELL SCHUBERT, AND G. W. WHELAND

CONTENTS

	PAGE
OCCURRENCE AND SIGNIFICANCE OF SEMIQUINONE RADICALS. By L. MICHAELIS	39
QUANTUM MECHANICAL BASIS OF THE STABILITY OF FREE RADICALS. By G. W. WHELAND.....	77
APPLICATION OF THE DROPPING MERCURY ELECTRODE FOR THE DETECTION OF INTERMEDIATE RADICALS. By OTTO H. MÜLLER.....	91
THE ANALOGY BETWEEN TWO-STEP OXIDATION AND TWO-STEP IONIZATION. By MAXWELL SCHUBERT.....	111
THE FREE ENERGY OF O_2^- IN RELATION TO THE SLOWNESS OF OXYGEN REACTIONS. By MANUEL H. GORIN.....	123
REMARKS ON THE ESTIMATION OF SEMIQUINONE RADICALS BY THE CONVERSION OF PARA-HYDROGEN. By L. FARKAS.....	129

* This series of papers is the result of a conference on Organic Free Radicals held by the Section of Physics and Chemistry of the New York Academy of Sciences, November 10 and 11, 1939. Manuscript received by the editor, February, 1940.

Publication made possible through a grant from the income of the Esther Herrman Fund, and the Ralph Winfred Tower Memorial Fund.

OCCURRENCE AND SIGNIFICANCE OF SEMIQUINONE RADICALS

BY L. MICHAELIS

From the Laboratories of the Rockefeller Institute for Medical Research, New York

STATEMENT OF THE PROBLEM

This review may be best started by propounding a few theses, then proceeding to the methods available for experimental verification, and to the discussion of the theoretical basis available for the explanation of the facts, and finally to the implications of the results for the kinetics of oxidation-reduction.

We may take as granted the definition of oxidation, as the detachment of one or more electrons. It depends on the particular properties of the substances involved whether or not the loss of the electron is accompanied by a simultaneous loss of a proton. If so, oxidation entails dehydrogenation. The essential thing in oxidation is the loss of the electron. In the same way, reduction is the acceptance of one or more electrons; which process is identical with hydrogenation only if the primary reaction product happens to be an acid of such strength as to add a proton under the prevailing conditions.

This review is concerned with *bivalent* oxidations and reductions. Thereby, we mean such processes in which not one, but two electrons are involved. Such bivalent oxidations or reductions are of the utmost importance in organic chemistry, because only for the case of bivalent oxidation can the constancy of valence of the elements C, H, O, N be maintained. Univalent oxidations produce free radicals. Such radicals have been known to exist for 40 years since Gomberg discovered triphenyl methyl. They are no longer considered as rarities. Yet, the number of free radicals capable of existence has been very considerably underrated indeed until a few years ago; the radicals known before were all, or practically all, of one particular type, namely those capable of existence when in the dissolved state, only in water-free organic solvents. That this property is not at all a general requisite for the existence of free organic radicals, will be shown in this review.

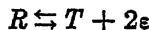
The thesis I propose as a starting point for the review is this: *every oxidation (or reduction) can proceed only in steps of univalent oxidations (or reductions).*

The question as to whether any bivalent oxidation can occur at all in a single step I dare not answer with certainty. There is for the

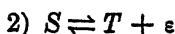
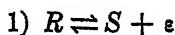
time being the alternative of two answers. Either one may concede the possibility of a bivalent oxidation in one single step, but has to attribute an extremely slow rate to such a reaction. Or one may say that a truly bivalent process in one step is the limiting case of more and more overlapping of two successive univalent steps, and that the utmost limit of overlapping, which means the attainment of a decidedly single bivalent step, may never strictly be realized.

The intermediate step in bivalent oxidation is a free radical, a molecular species with an odd number of electrons, designated by G. N. Lewis as an "odd molecule." We shall show many cases in which a free radical can exist in a true, thermodynamical equilibrium with its oxidation and its reduction product and is, to the extent of its thermodynamically defined equilibrium concentration, not a fragile molecule, but just as stable in time as ordinary organic compounds. It should be stated with emphasis that all radicals, unless in the solid state, establish this equilibrium with unmeasurable speed, whereas the over-all bivalent oxidation or reduction of organic compounds is known to be in many cases a sluggish process, characterized by a high activation energy. Free radicals, in solution, can exist only in equilibrium with what we may call their two parent substances, and usually this equilibrium is quite in favor of the parent substances. This fact often makes the recognition of the radicals somewhat difficult. It has been responsible for the rather astounding fact that the vast majority of those free radicals to be dealt with in this review, had not been known at all only a few years ago.

The equilibrium mentioned may be described as follows. In a bivalent oxidation-reduction system we may distinguish three different levels of oxidation. We have the reduced form, R ; the intermediate (semi-oxidized) form S , which is a free radical, and the totally oxidized form, T . The overall process of the bivalent process may be written



where e is the electron. This process, if separated into two steps, is



The equilibrium is ruled by the reversible process



The equilibrium constant therefore is

$$\frac{[S]^2}{[R] \cdot [T]} = k = \frac{1}{\kappa}$$

We shall refer to k as the *radical formation constant*, or *semiquinone formation constant*; and to its reciprocal, κ , as the *dismutation constant* because the process in the direction



is adequately designated as dismutation (or disproportionation). (However, those dismutations previously studied are bivalent in either direction, such as the Cannizzaro reaction; they are all sluggish reactions.)

A free radical may be capable of chemical reactions other than dismutation. In the first place, it may be capable of forming a valence saturated compound of double molecular size, or a dimer, D . The equilibrium of the process



is characterized by the dimerization constant q

$$\frac{[D]}{[S]^2} = q$$

It should be emphasized that the process of dismutation always occurs with unmeasurable speed, in the same way as does an ionization, whereas the dimerization often proceeds at a slower, measurable rate; for instance, the dimerization of triphenylmethyl, and the dissociation of hexaphenylethan into two radicals, is not at all instantaneous. The high speed of dismutation or its reversal is in striking contrast to the sluggishness of the majority of chemical reactions in organic chemistry, except for ionizations which also are instantaneous.

In what follows we shall consider the existence of free radicals only to such an extent as compatible with thermodynamical equilibrium, and disregard entirely their existence in excess of equilibrium conditions. Radicals can be produced by some artifices at a concentration by far exceeding their equilibrium conditions. In this case they usually have an extremely short life time. However, we shall never resort to any concentration of a radical in the dissolved state in excess of its equilibrium concentration. Even, as we shall proceed to kinetic considerations, we shall maintain this principle, and in dealing with kinetics, we shall never resort to conditions other than thermodynamically defined, or at least nearly defined, at each instance of time.

All oxidation-reduction systems in which the radical formation constant is so large that an easily recognizable concentration of the

radical can be formed, are reversible. The reversed problem, as to whether all reversible systems are those with an appreciably high radical formation constant, will also be answered in the affirmative in the following discussions. The value of the radical formation constant k may vary greatly even for one substance at varied pH's. The highest values of k encountered were about 1000. There is no definite lower limit, but the methods now available are not capable to distinguish a constant < 0.01 from 0. This insensitivity of the methods has retarded the appreciation of free radicals for many years.

METHODS FOR QUALITATIVE AND QUANTITATIVE DETERMINATION OF FREE RADICALS

There are three methods available for the detection and quantitative estimations of free radicals: the potentiometric method which is the most powerful and most diversified of all; the measurement of magnetic susceptibility, which seems to have been the most convincing one for the sceptics among the organic chemists; and finally optical methods may be useful in suitable cases. Dr. Müller, in a special paper, will discuss a fourth method, the polarographic method.

The Potentiometric Method

In a reversible oxidation system the potential E at a noble metal electrode, for a univalent system (such as ferricyanide + ferrocyanide) varies during a reductive or oxidative titration as follows

$$E - E_m = \frac{RT}{F} \ln \frac{\text{oxidized form}}{\text{[reduced form]}} = 0.06 \log_{10} \frac{\% \text{ oxidation}}{100 - \% \text{ ox.}}$$

where E_m is a constant, namely the potential in the midpoint of titration, in volts, at 50% reduction; usually designated as the *normal potential*. The value of this constant may depend on pH. In this case, the normal potential at a given pH is designated by Clark as E_h , and the value of E_h holding for pH = 0 is designated as E_0 . The value $RT/F \times \log_{10} e$, at 30° C. is 0.06001, in what follows abbreviated to 0.06 (e. g. in formula 10).

In a bivalent oxidation, provided there is no intermediate univalent step, the same formula holds replacing only the factor RT/F by $RT/2F$. The definitions of E_m , E_h , E_0 , are the same as before. When the potential is plotted against the degree of oxidation, one obtains a sigmoid curve symmetric around its midpoint at 50% oxidation. In order to accentuate this symmetry and hereby facilitate the mathe-

matical analysis to follow, we take as zero point of the abscissa of the titration curve, this midpoint, at 50% oxidation. So we count the starting point of the titration as - 1 and the endpoint as + 1. This scale of the abscissa will be referred to as the μ -scale. In terms of it, the electrode equation of a bivalent oxidation without any intermediate steps is this:

$$E - E_m = \frac{RT}{2F} \log \frac{1+\mu}{1-\mu} \quad (1)$$

showing the symmetry of the potential around $\mu = 0$. We shall show, that this shape of the equation is, though very often nearly, yet never perfectly obeyed, on account of an intermediate step occurring.

On taking into consideration the formation of the intermediate radical also, the strict equation of the potential, as dependent on the degree of oxidation-reduction can be derived from the following set of equations:

At any point of the titration we have

$$r + s + t = a \quad (2)$$

where r, s, t are the molar amounts of R, S, T ; and a that of the dye in all its forms. Furthermore, on performing the titration of the dye in its reduced form, R , with an oxidizing agent, we have

$$s + 2t = x \quad (3)$$

where x is the equivalent amount of the oxidizing agent added. Furthermore, the equilibrium of r, s , and t is determined by

$$\frac{[s]^2}{[r] \cdot [t]} = k \quad (4)$$

where k depends on pH, since S, R , and T may change their state of acidic ionization with change of pH.

These three equations (2, 3, 4) can be solved for r, s , and t in terms of a, x , and k . Having r, s , and t , we put their values into the potential equation

$$E = E_m + \frac{RT}{2F} \ln \frac{t}{r} = E_1 + \frac{RT}{F} \ln \frac{s}{r} = E_2 + \frac{RT}{F} \ln \frac{t}{s} \quad (5)$$

where E_m, E_1 and E_2 are constants such that always

$$E_m = \frac{E_1 + E_2}{2} \quad (6)$$

By a suitable rearrangement, and using the μ -scale instead of x , and defining

$$\gamma = 4\kappa - 1 = \frac{4 - k}{k} \quad (7)$$

the most useful form of the potential equation is obtained as follows:

$$E - E_m = \frac{RT}{2F} \ln \frac{1 + \mu}{1 - \mu} + \frac{RT}{2F} \ln \frac{\sqrt{1 + \gamma(1 - \mu^2)} + \mu}{\sqrt{1 + \gamma(1 - \mu^2)} - \mu} \quad (8)$$

showing symmetry around $\mu = 0$. The semiquinone formation constant appears only in the second logarithmic term, and only in the form γ . As k becomes very small, γ becomes very large, and the second logarithmic term approaches the value $\ln \frac{\sqrt{\gamma}}{\sqrt{\gamma} - \mu}$ and so approaches zero. In practice it seems never to vanish entirely.

An interesting singular case arises when $k = 4$, and so $\gamma = 0$. Then the second logarithmic term becomes equal to the first, and we obtain, for $k = 4$:

$$E - E_m = \frac{RT}{F} \ln \frac{1 + \mu}{1 - \mu}$$

with the factor RT/F instead of $RT/2F$; as though it were a univalent oxidation. Since k for one particular dye stuff can be varied by varying pH, it is easy to make $k = 4$ for many dyestuffs by a proper choice of pH. The occurrence of such a curve will never cause any confusion with a real univalent curve, because during the titration there appears the color of the intermediate radical.

FIGURE 1 shows a family of curves for various values of k . They have only one point of inflection, at $\mu = 0$, when $k \leq 16$. They have three points of inflection when $k > 16$. The two lateral points of inflection lie at those two points where

$$\mu = \pm \frac{1}{2} \sqrt{\frac{k - 16}{k - 4}} \quad (9)$$

In order that this equation may have a physical significance, the value of μ must lie between -1 and $+1$. This is the case only when $k > 16$. For, if $k < 16$, μ is an imaginary number; and when $k < 4$, μ is, although real, yet > 1 , and has no physical significance. The result is that *there can exist lateral points of inflection only when $k > 16$* .

The potential difference (always taken as a positive number) between $\mu = 0$ and $\mu = \pm 0.5$ (that is, between 50% oxidation, and either 25% or 75% oxidation) will be designated as E_i , the *index potential*. Its graphic evaluation from any experimental titration curve is an easy task giving valuable information before trying to

interpret the curve as a whole. From E_i (in volts) it is easy to arrive at k , by the equation

$$k = \left[10^{\frac{E_i}{0.06}} - 3 \cdot 10^{-\frac{E_i}{0.06}} \right]^2 \quad (10)$$

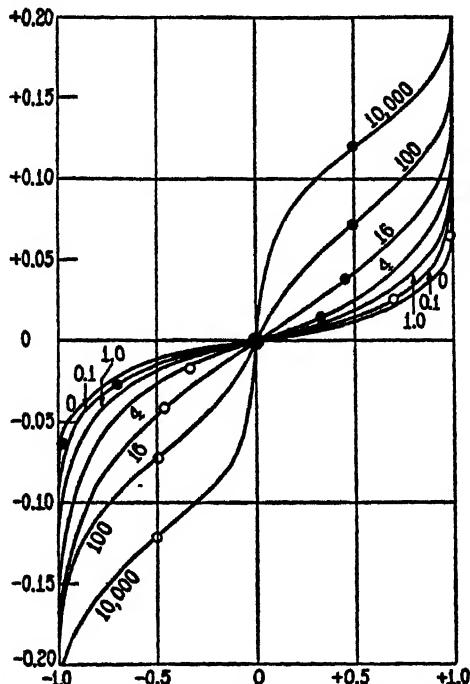


FIGURE 1. Two-step titration curves involving formation of semiquinone radical. The degree of oxidation, expressed in the μ scale: $\mu = 0$ means 50 per cent of the total oxidation; $\mu = 1$ means 100 per cent of the total oxidation. Ordinates: $E - E_m$, the potential, referred to the mean normal potential E_m , in volts. Each curve holds for the value of k (semiquinone-formation constant) as indicated. White circle: that point of the titration curve where $E = E_1$ (the normal potential of the lower step). Black circle: that point where $E = E_2$ (the normal potential of the higher step). The black and white circle in the center belongs to the curve for $k = 1$; here E_m , E_1 , and E_2 coincide at $\mu = 0$. White circles are on the left side for curves with $k > 1$; they are on the right side when $k < 1$. Both the white and the black circle in the curve for $k = 0$ belong actually to a curve for k intermediate between 0 and 0.1. When k is precisely 0, the circles would lie at $\mu = \pm 1$, and at potential $\pm \infty$. The lateral points of inflection begin to appear only when $k > 16$.

That point of the curve, where $r = t$, is of course at $\mu = 0$. Its potential is called the mean normal potential, E_m . That point where $r = s$, may be called the normal potential of the lower step of oxidation, or E_1 ; where $s = t$, we have the normal potential of the higher step, E_2 . When $E_1 < E_m < E_2$, we speak of the *natural order of the three normal potentials*. It occurs whenever $k > 1$. When $E_1 > E_m >$

E_2 , we speak of the *reversed order of the three normal potentials*. It occurs when $k < 1$. When $k = 1$, $E_1 = E_m = E_2$.

The potential, during any one oxidative titration, starts from $-\infty$; at some point it equals E_1 , at some other point it equals E_m , at another E_2 , and finally reaches $+\infty$. E_m , of course, is the potential at $\mu = 0$. The potential equals E_1 , when

$$\mu = -\frac{k-1}{2k+1} \quad \left. \right\} \quad (11)$$

and it equals E_2 , when

$$\mu = +\frac{k-1}{2k+1}$$

This shows, that for very large k the two μ -values are ± 0.5 . In this case, the curve is entirely separated into two successive curves each for a univalent system. Furthermore it shows that for $k > 1$ there is the natural order of the three normal potentials, and for $k < 1$ the reversed order, as stated before.

One of the most important problems is to calculate the maximum fraction of the dyestuff which can exist in the form of the semiquinone radical. It is easy to see that the maximum ratio of semiquinone to total dye, $(s/a)_{\max.}$, obtains in the midpoint of titration, where $\mu = 0$, when the dye is reduced to 50%. The following equation correlates this maximum with the semiquinone formation constant k :

$$(s/a)_{\max.} = \frac{\sqrt{k}}{2 + \sqrt{k}} \quad (12)$$

The following table correlates k with $(s/a)_{\max.}$:

k	$(s/a)_{\max.}$
0.0001	= 0.0050
0.001	= 0.0156
0.01	= 0.0476
0.1	= 0.137
1	= 0.333
10	= 0.61
100	= 0.83
1,000	= 0.95
∞	= 1.00

Herefrom it can be seen that e. g. even for a k as small as 0.001, still 1½% of the dye can exist as semiquinone. Now, the present methods are not able to distinguish a $k = 0.001$ from a $k = 0$; not even a $k = 0.01$ can be safely distinguished from 0. Therefore, potentiometric

titration can reveal the existence of a semiquinone only when k is > 0.01 . Yet, the potentiometric method is at the present time the most sensitive one for the purpose. This is why semiquinone radicals have escaped observation until recently.

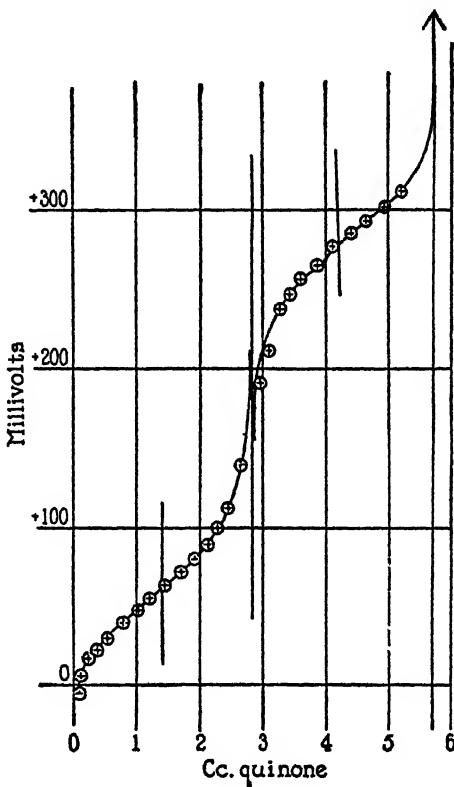


FIGURE 2. Oxidative titration curve for α -oxyphenazine. The leuco-dye is titrated with benzoquinone at pH = 1.00. Abscissa: quinone added. Ordinate: potential, in millivolts, referred to the normal hydrogen electrode. Since $E_i > 40$ mv. (namely 75 mv.), there are three points of inflection.

Coming back for a moment to equation (9), concerned with the lateral points of inflection, it has been found that these lateral points exist only when $k > 16$. Since the index potential E_i is correlated to k in the manner expressed in equation (10), one may ask the question, what E_i corresponds to $k = 16$? The answer is: 40 millivolts. Hence, *there are no lateral points of inflection unless $E_i > 40$ mv.*

How these things look in practice, some diagrams will show. FIGURES 2 and 3 show some individual potentiometric titration curves with two

more or less overlapping steps. For each dyestuff a set of such titration experiments is performed, each at constant pH, showing various forms of the curve, from each of which the particular E_1 can be read and the values of E_1 , E_2 and k can be computed. When these values, obtained from a set of experiments at varied pH, are plotted against pH, one obtains curves of which FIGURES 4 to 7, and 10 are examples.

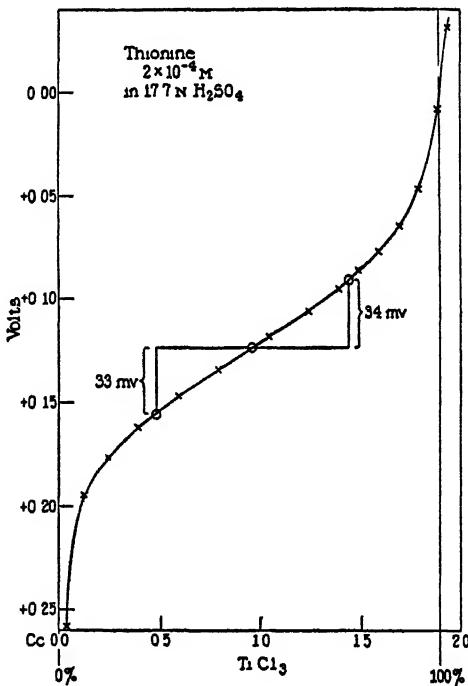


FIGURE 3. Reductive titration curve for thionine in 17.7 N H₂SO₄. The dye is titrated with TiCl₃. Abscissa: TiCl₃ added. Ordinate Potential in volts, referred to an arbitrary zero point. Since $E_1 < 40$ (namely 33 m.v.), there is only one point of inflection.

Each curve, in general, consists of linear sections of various slopes, connected by slightly curved transition zones. Whenever there is a point of inflection it indicates that here the pH equals the pK ($-\log$ of an acidic ionization constant) of one of the molecular species involved. The following rules established by W. M. Clark may be applied for interpretation of these bends:

- 1) All slopes which can occur in a univalent system (such as are represented by the E_1 or the E_2 curves) are either 0, or 0.06, or 0.12 or 0.18 volt per pH unit.

2) In a bivalent system (the E_m curve) there may occur slopes of 0; 0.03; 0.06; 0.09 volt per pH unit.

3) Each single ionization constant changes the slope by 0.06 volt per pH unit, if the system is univalent (E_1 and E_2); it changes the slope by 0.03 volt per pH unit, if the system is bivalent (E_m).

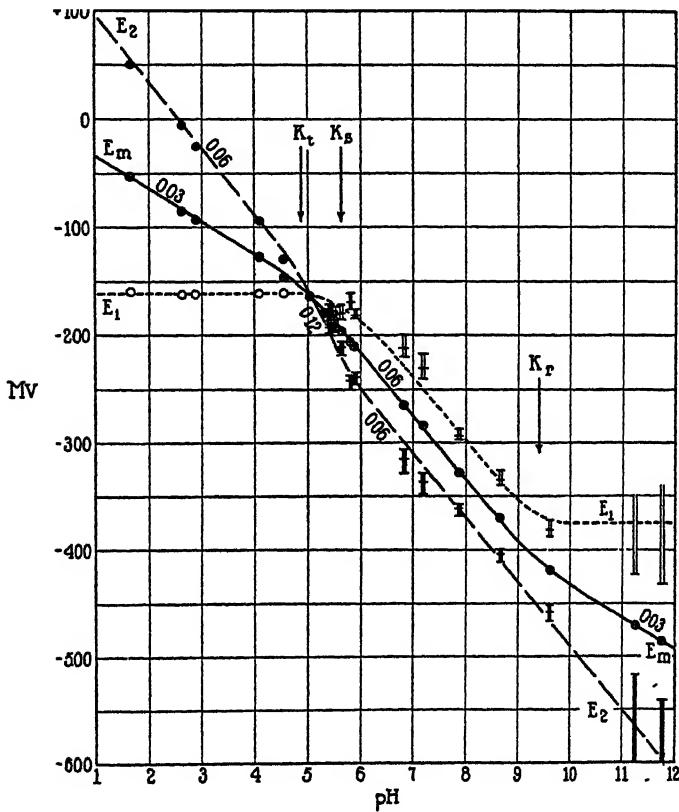


FIGURE 4. The three normal potentials E_1 , E_m , E_2 , of pyocyanine plotted against pH. K_t , K_s , K_r , acidic ionization constants of the oxidized, semiquinone, and reduced forms. (The abscissa is the corresponding pK). The slopes of the rectilinear parts are either 0, or 0.03, or 0.06 or 0.12 volts per pH unit. At the crossing point $E_1 = E_m = E_2$, and the semiquinone formation constant $k = 1$. To the left hereof, there is the natural order of the three normal potentials, and $k > 1$; to the right, there is the reversed order, and $k < 1$.

4) If the ionization constant belongs to the reduced form of the particular system, the bend consists in a flattening; if it belongs to the oxidized form, it consists in a steepening.

In general it can be seen from these examples that strong separation of the steps, such that the natural order of the three normal potential

occurs, takes place for cationic semiquinones only in strongly acid solutions, for anionic semiquinones only in strongly alkaline solutions. Why this should be so will be discussed in a later section.

There is another example of the same type, discovered quite recently, in which also the semiquinone formation constant largely depends on acidity. This is for thiazine dyes, such as thionine (FIGURE 8). The range of acidity in which the separation into two steps becomes distinct is so high that it can no longer be expressed in terms of pH, because here definition or measurement of pH breaks down; namely in 10 to 26 N H₂SO₄. At any rate, FIGURE 8 shows that the separation of the two steps becomes greater as acidity increases. For methylene

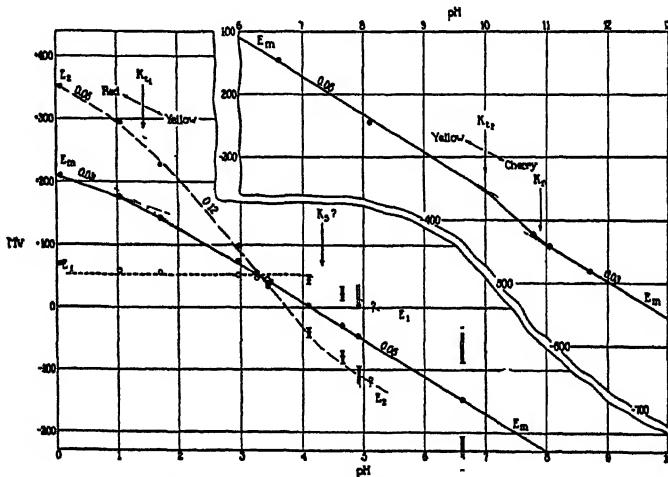


FIGURE 5. The same as FIGURE 4, for α -oxyphenazine.

blue (FIGURE 9), there are two steps in strongly acid solutions too; the separation of the two steps is not sufficient even in 23 N H₂SO₄ to make any lateral points of inflection to appear.

All this is valid provided the intermediate form is a free radical, which is a molecule of the same molecular size as either of the two parent substances. In order to prove that this is the case, one has only to carry out two titration experiments at the same conditions of pH, but with varied initial concentration of the dye. If the intermediate form is a radical, the curves must coincide (FIGURE 11, 12). If the intermediate form is a dimer the shape of the curve depends very largely on the initial concentration, because a bimolecular reaction is involved in establishing the equilibrium. The following FIGURES (13, 14, 15) show examples where the shape of the titration

curve does depend on the concentration of the dye, indicating that the intermediate step of reduction is not, or not alone, the radical. On decreasing the concentration one may reach a concentration range in which the shape becomes independent of concentration. Within this range, the intermediate form is practically entirely represented by the radical. Here, the radical formation constant can be calculated.

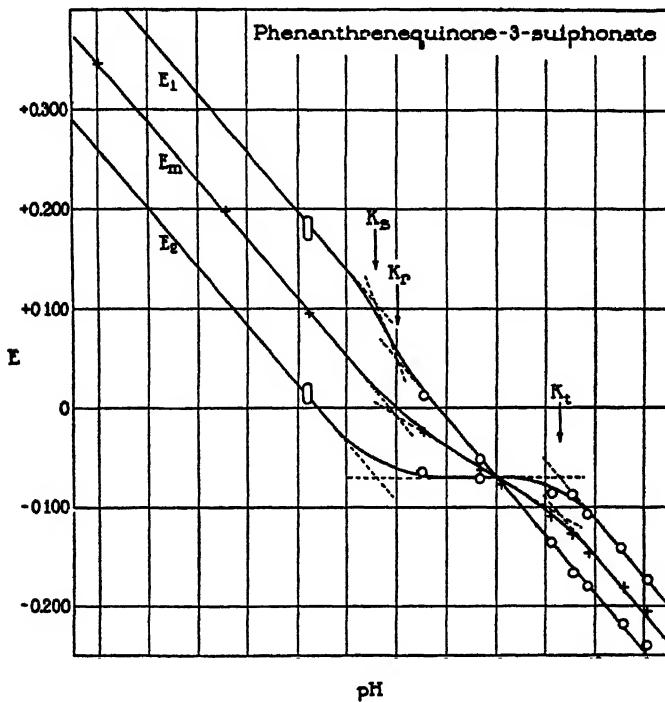
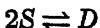


FIGURE 6. The same, for phenanthrene quinone-3-sulfonate. Here, the natural order of the three normal potentials is on the right hand side of the crossing point, the reversed order on the left hand side.

Having this, the change of the curve at higher concentration allows to calculate the equilibrium constant of dimerization



The best studied case in this respect is phenanthrenequinone sulfonate. In an alkaline solution, at high dilution much of the radical is formed during the reduction. At higher concentration, the dimeric form of the intermediate step is also formed. In an acid solution, the free radical is not, or scarcely at all, detectable, and at low concentration

no dimeric form is either. At higher concentration, much of the dimeric intermediate form is capable of existence, but not the free radical to any measurable extent. FIGURES 16 and 17 show the percentage at which the two intermediate forms can exist in maximo, one in an acid solution, the other in an alkaline solution.

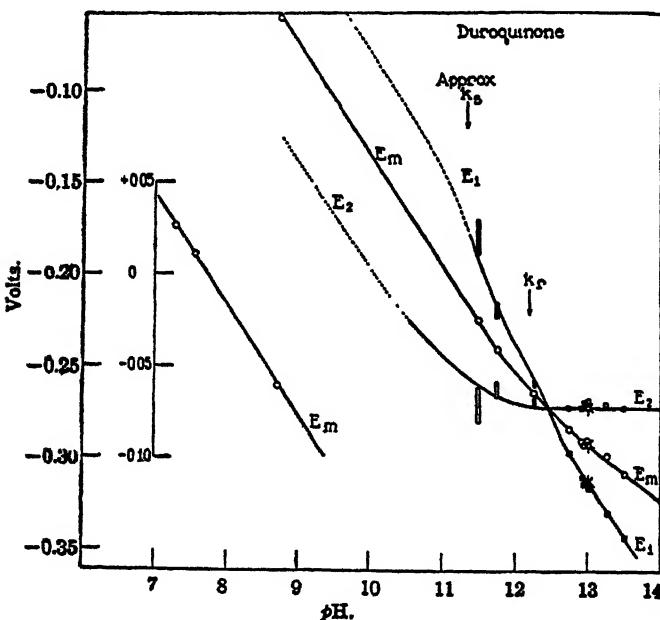


FIGURE 7. The same, for duroquinone.

The Magnetometric Method

Since in such organic compounds as we are considering here, atoms with incompletely filled inner shells never occur, it is safe to follow G. N. Lewis' assumption that all valence-saturated compounds, with an even number of electrons should be diamagnetic, and radicals, with one unpaired electron, should be paramagnetic. In polyatomic organic molecules all orbital permanent magnetic moments are quenched, and paramagnetism can arise only from the spin of an unpaired electron. The paramagnetic susceptibility (disregarding the diamagnetism), according to van Vleck, should be, per mole of a free radical

$$\chi_m = \frac{4N\beta^2S(S+1)}{3KT}$$

Here β is the Bohr magneton, N is Avogadro's number, K is Boltzmann's constant, T the absolute temperature, S the spin quantum number of the electron, always = $\frac{1}{2}$ for an ordinary radical with only one unpaired electron. This gives, at 20° C. per mole of a radical a susceptibility of 1260×10^{-10} units; and *vice versa*, the susceptibility obtained in a substance, even in the dissolved state, allows to calculate

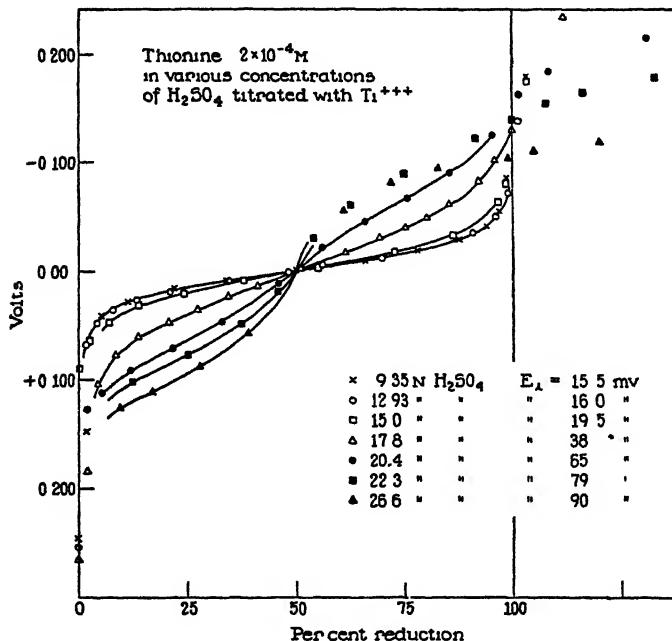


FIGURE 8. A family of reductive titration curves for thionine, in various solutions of H_2SO_4 . Abscissa: Titanous chloride added, in per cent of the total equivalent. Ordinate: Potential, for each single curve referred to its potential at 50 % reduction. For the two highest concentrations of H_2SO_4 , the potentials in the second half of the curves overlap somewhat with those of the Ti^{++++}/Ti^{+++} system. All drawn out curves are graphic interpolations between such experimental points not affected by overlapping.

Curves with an index potential $E_i < 40$ mv. show one point of inflection, those with $E_i > 40$ mv. show three points of inflection.

the concentration of the radical and to compare it with the value obtained by the potentiometric method. The difficulty, however, in doing so, is obvious. In a dilute solution, the diamagnetism of the solvent is very much greater than the paramagnetism of the radical in its high dilution. Furthermore it is scarcely possible to establish and to maintain any definitely known degree of reduction of a dye solution in the container of the magnetic balance. The case is much more difficult than for crystals and for systems not sensitive to oxygen, and

especially difficult for the investigation of a solution, containing the paramagnetic substance in equilibrium with its diamagnetic parent substances. The difficulty was overcome in the following way.

The quinonoid form of the dye, dissolved in a suitable solvent, is mixed with a slow-reducing agent, such as to stretch the period of reduction over one to two hours, and the change of magnetic susceptibility is measured during this period. The diamagnetism prevailing at the beginning will decrease in time and then drift back so as to attain a time-independent value, showing the development, and later the disappearance of a paramagnetic molecular species. This method shows directly the increment of magnetic susceptibility due to the free radical at any instant, and the change of this increment in time. No correction for the diamagnetic susceptibility arises, since the change in diamagnetic susceptibility of the dyestuff due to reduction, and due to the oxidation of a properly chosen reducing agent, is perfectly negligible; a fact which makes unnecessary all corrections for diamagnetism according to Pascal. The paramagnetic increment at the time of its maximum can be compared with data obtained potentiometrically, since the maximum point corresponds to the half reduced state, and potentiometric methods allow to calculate the ratio of semiquinone to total dye in the half reduced state.

So far, two slow-reducing agents, working in homogeneous solution have been found suitable. In alkaline solution, glucose can be used.¹ According to its concentration and to pH, the rate of reduction can be properly controlled. In slightly acid solution, methylglyoxal, in presence of KCN as a catalyst, is a slow-acting reducing agent.² The rate of reduction can be regulated by a suitable concentration of the catalyst CN⁻.

Purely qualitatively the transient appearance of a radical during the reduction can be demonstrated by this method for many dyestuffs. The viologens, or phenanthrenequinone sulfonate (FIGURE 18) in alkaline solution are suitable examples. The latter substance, when being reduced in an acid solution, shows no change in magnetic susceptibility, demonstrating that in this case the brown intermediate form of reduction is not a radical. In order to evaluate this method quantitatively one has to select a substance which easily forms the free radical, but never forms any dimerization product of the radical. Such a case is represented by duroquinone, which has the desired properties due to its particular structural configuration, as will be shown later.

¹ Michaelis, L., Boeker, G. F., & Reber, R. K. Jour. Am. Chem. Soc. 60: 202. 1938.

² Michaelis, L., Boeker, G. F., & Kuck, J. A. Jour. Am. Chem. Soc. 60: 214. 1938.

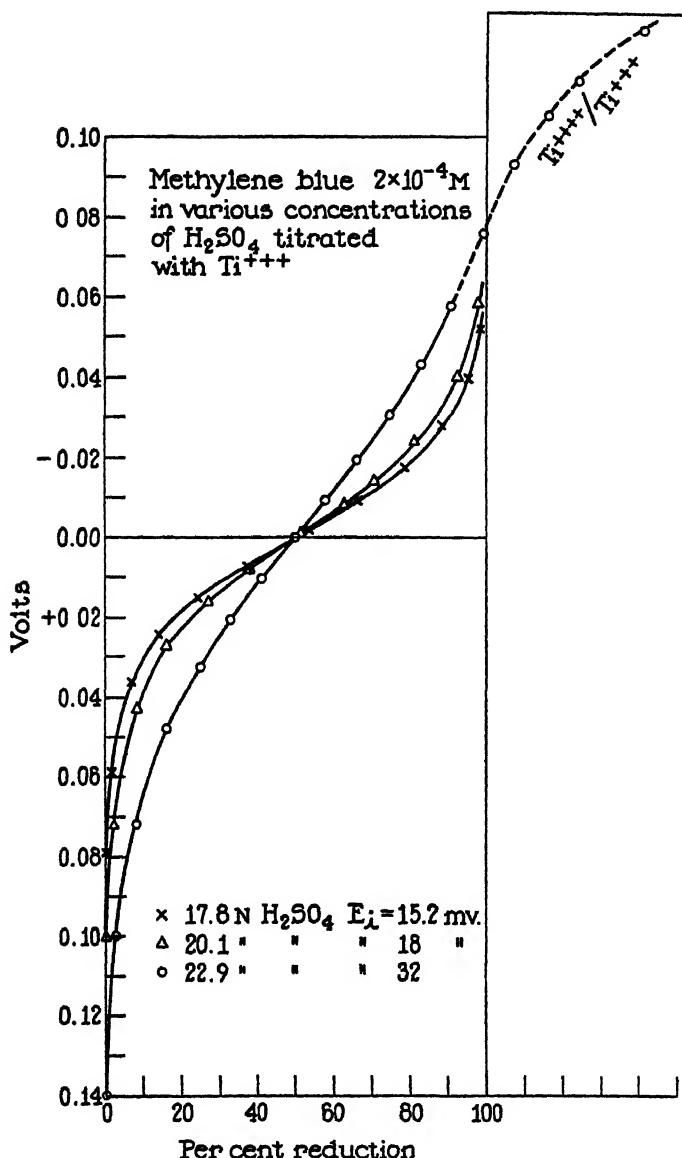


FIGURE 9. The same as FIGURE 8, for methylene blue. The index potentials, E_i , in the whole range of acidity covered by the experiments, never reaches 40 mv., hence there is always only one point of inflection.

The quantitative evaluation of the concentration of the radical of duroquinone in the half-reduced state by this magnetic method agrees very satisfactorily with that obtained potentiometrically at the same pH (FIGURE 19). The ratio of semiquinone to total dye at pH = 13.0 was found to be 0.52, both magnetically and potentiometrically. The magnetic method confirms the results of the potentiometric one. I believe that the reluctance of renowned organic chemists in accepting the idea of the existence of these radicals has been overcome only by these convincing magnetic measurements. Unfortunately, this

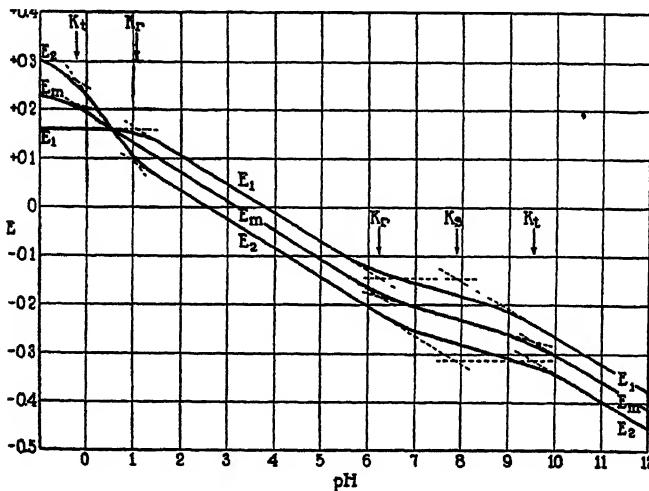


FIGURE 10. The three normal potentials of riboflavin plotted against pH. Only in extremely acid solution, there is the natural order of the three normal potentials.

method is not sensitive enough, and requires the concurrence of many suitable properties of the substance, so it cannot compete with the potentiometric method in general practical applicability.

Optical Methods

All semiquinone radicals are strongly colored. Since they are, when in solution, in equilibrium with the quinonoid form of the dye and the leucodye the strong color of the quinonoid form may more or less overshadow the color of the radical according to the ratio of concentrations existing in this equilibrium state. If the overlapping of the steps is not too great, the pure color of the radical appears in a suitable stage of reduction. In quite a number of cases, *e. g.* for several quinones, for aromatic diamines, for the viologens, and for the system

benzil-benzoin, the quinonoid form is so lightly colored that the intensely colored radical is easily visible even in high dilution. Most radicals exhibit a distinct set of sharp absorption bands. However, as yet, it is impossible to infer from the absorption spectrum whether or not it represents a free radical.

Examples of absorption spectra of some semiquinone radicals are shown in FIGURES 19 and 20.

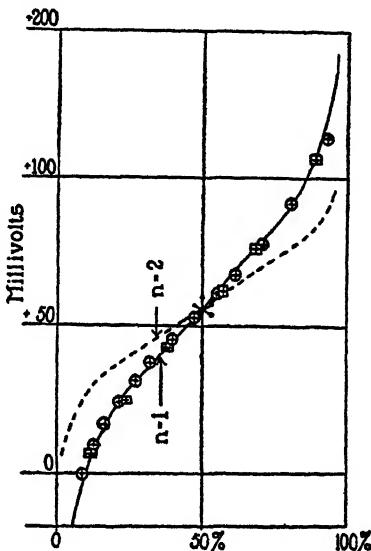


FIGURE 11. A case where the potential is independent of the initial concentration of the dye showing that the intermediate form is a free radical. Solvent: 0.05 M HCl + 0.1 M KCl solution. Lenco-a-oxyphenazine at pH 1.38 is titrated with quinone. The two steps are widely separated, and only the first step of the oxidation is shown in this diagram. The first titration is represented by the circles. After finishing the experiment, the solution was diluted with the above solvent to a three-fold volume, the dye re-reduced by H₂ + palladium, the hydrogen expelled with N₂, and the titration with quinone was repeated. It furnished the points marked with squares. The drawn out line is the one calculated for a univalent oxidation (electron number $n = 1$).

Sometimes, a colorimetric method can be used in identifying a colored substance as a free radical. A good example is this. On oxidizing benzil, C₆H₅ · CO · CHOH · C₆H₅, to benzine C₆H₅ · CO · CO · C₆H₅, an intense purple color arises as an intermediate step of the oxidation, provided one works in strongly alkaline solution. By colorimetrically measuring the amount of colored material on varying only the volume of the solvent, it was found that this amount is independent of the volume.⁸ This is possible only if the intermediate

⁸ Michaelis, L., & Fetscher, E. S. Jour. Am. Chem. Soc. 59: 1246. 1937.

substance has the same molecular size as each parent substance, and no dimeric compound is formed. This shows that the purple substance is the free radical, which may be written, preliminarily, $C_6H_5 \cdot CO \cdot COH \cdot C_6H_5$, with one trivalent C atom, with reservation as to discuss this formula presently once more.

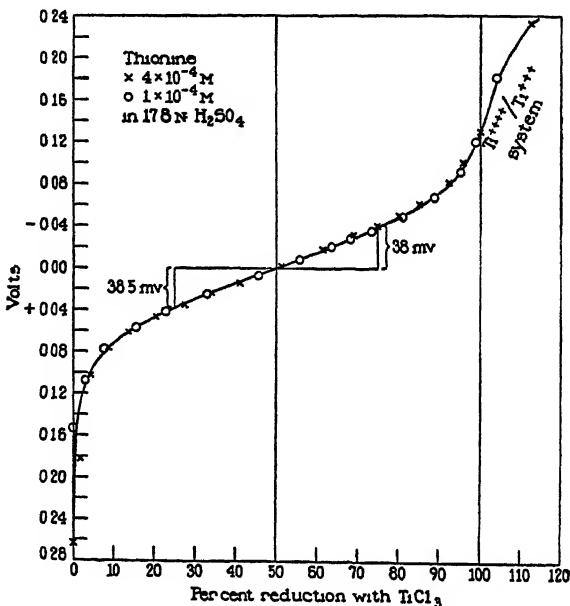


FIGURE 12. Another case where the shape of the titration curve is independent of the initial concentration of the dye, indicating that the intermediate form is a free radical. Reductive titration of thionine in 17.8 N H_2SO_4 .

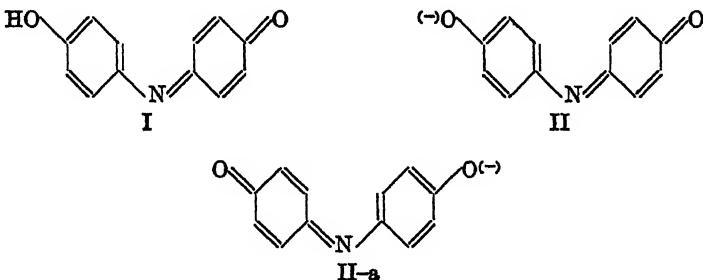
THE THEORY OF RESONANCE

Emphasizing once more that the equilibrium between a radical and its products of dismutation is always instantaneously established, one cannot expect any appreciable concentration of a radical in the dissolved state except for the case that the radical is comparable in stability with its parent substances. In the crystalline state, this need not be so. Whether or not the free radical can be prepared in the solid state from such a solution, depends on solubility properties. There is a sufficient number of cases where a solid radical, exhibiting the full expected paramagnetic solubility, can be prepared. The fact that such an occurrence is relatively rare, is due to the tendency of most of the radicals to dimerize when present in high concentration.

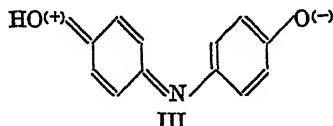
The solid state is certainly the highest concentration of the substance obtainable. Only such radicals with no tendency of dimerization can be expected to exist in the crystalline state as free radicals. We shall discuss later what kind of structure is favorable to the prevention of dimerization.

First of all, however, it should be discussed, what kind of structure is favorable to a high stability of a radical. The answer is: stability is favored by *equivalent resonance* of a special type, which may be designated as the *semiquinone resonance*.

By equivalent resonance, we understand resonance between two (or more) limiting structures which are equivalent. We have to consider two quite different types of equivalent resonance in dyestuffs. One type occurs in the regular, quinonoid forms of the dyestuffs. It may be illustrated by an example. Phenolindophenol (I), in alkaline solution, has the structure II.

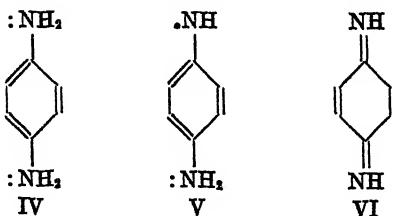


This formula II can be shifted to II-a, differing from II only by the orientation of the molecule, as a whole, in space. Electrons only need be shifted, while the nuclei are kept in their place. In II, the left hand oxygen is negatively charged, the left hand ring is benzenoid, the right hand ring is quinonoid. In II-a, the mirror image structure for II, right and left is exchanged, but except for orientation in space, it is indistinguishable from II itself. Consequently, the energy content of II is the same as that of its mirror image, II-a, therefore there will be a resonance structure, to which these two limiting structures will contribute an equal share. The real structure cannot be expressed by any single formula. Since each limiting structure contributes an equal share, we have the most favorable condition for stabilizing the molecule by resonance. In I, resonance is no longer equivalent. Shifting electrons does not produce the mirror image of I, but a structure such as III which has less probability of existence than I; resonance between I and III is not equivalent, but more in favor of I. The

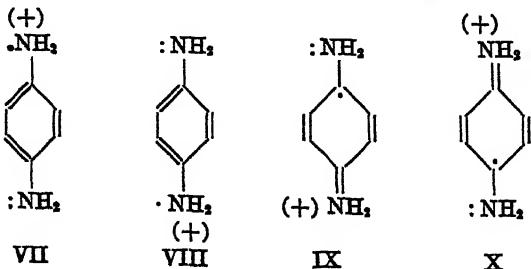


stabilizing effect of non-equivalent resonance is usually smaller than that of an equivalent one. It is characteristic of the resonance in quinonoid dyestuffs that there is an ambiguity as to which of two (or more) rings is quinonoid, and which is benzenoid. This ambiguity makes the "quinonoid" dyestuffs very much more stable than the simple quinones themselves, having only one ring of definitely quinonoid structure.

These very perfunctory remarks on the quinonoid dyestuffs have to suffice, and we proceed now to the other type of symmetric resonance as exhibited by the semiquinone radicals. Again, an example will illustrate the matter. Para-phenylene-diamine (IV) can be oxidized



by a univalent oxidation to the free radical V, and by a bivalent oxidation to the quinonoid structure VI, quinone diimine. The radical V is under a certain condition much more stable than the diimine. This condition is that it should be in a moderately acid solution. In this case, it attaches a proton and now has the structure VII. This is in equivalent resonance with VIII. Beside this pair of limiting structures, other pairs of structures with equivalent resonance may be imagined, by placing the odd electron somewhere else. One such pair of such structures is IX and X. VII and VIII are benzenoid, IX and



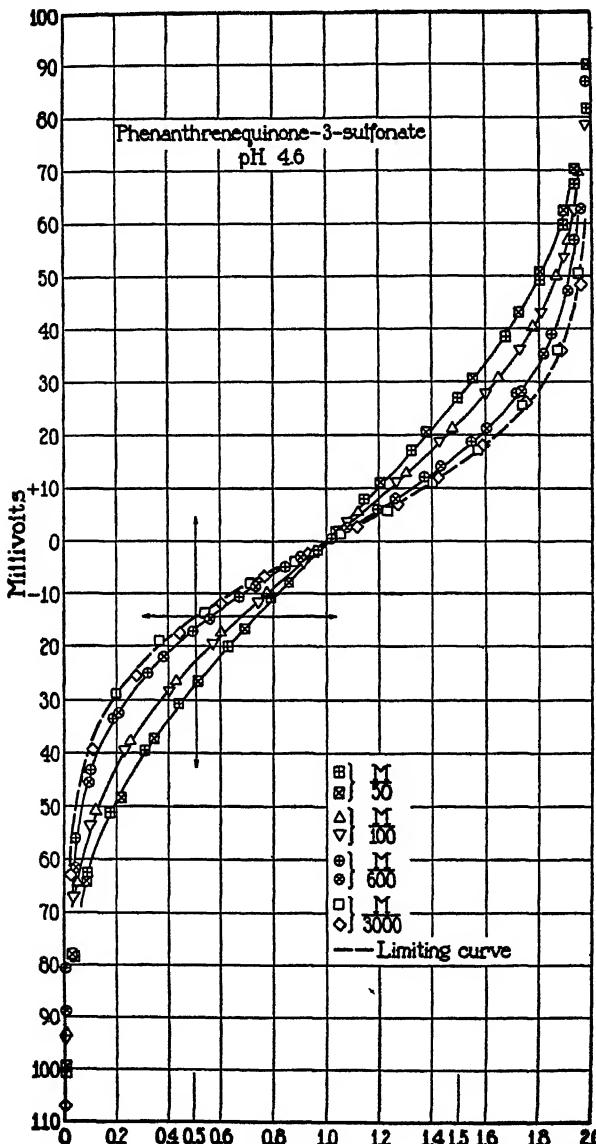


FIGURE 13. Titration curves for phenanthrene quinone-3-sulfonate at pH 4.6, with varied initial concentrations of the dye-stuff. Abscissa is the amount oxidizing agent added, in equivalents, such that 2.0 is the same as 100% of the total oxidation.

The limiting curve for infinitely low concentration shows an $E_{1/2}$ almost indistinguishable from 14.3, indicating that the intermediate form does not exist to any appreciable percentage in very low concentration and that the intermediate form such as exists in higher concentrations is the dimer, not the free radical.

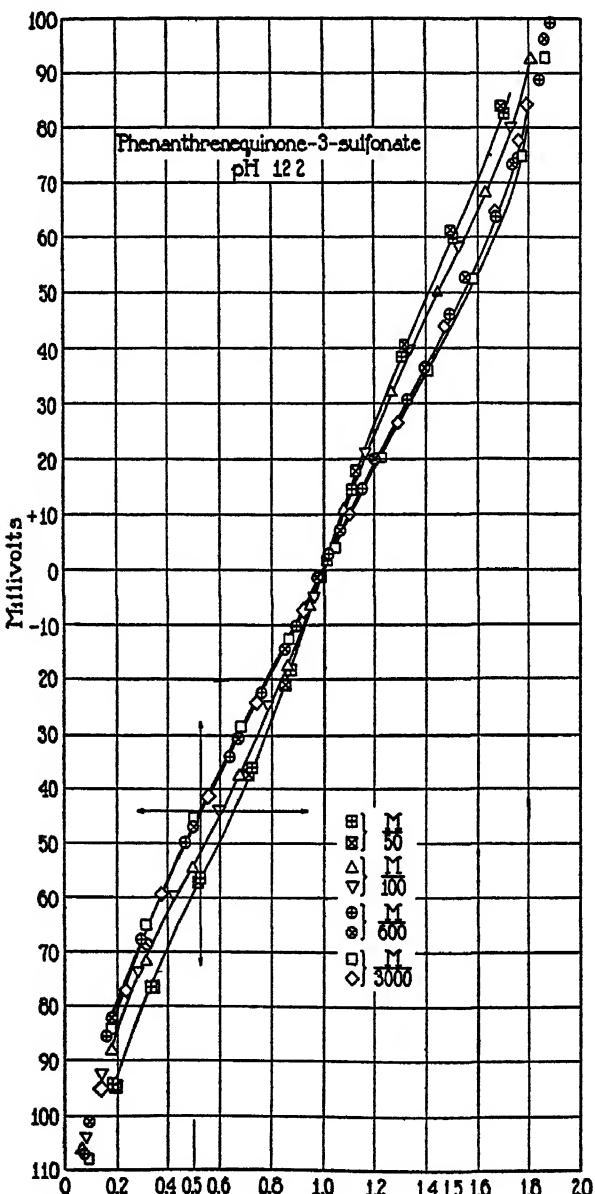


FIGURE 14. The same as FIGURE 13, at pH 12.2. Here the limiting curve for infinitely low concentration has an index potential of 48 mv. indicating that there is a free radical. In higher concentration, the radical is in equilibrium with a dimer.

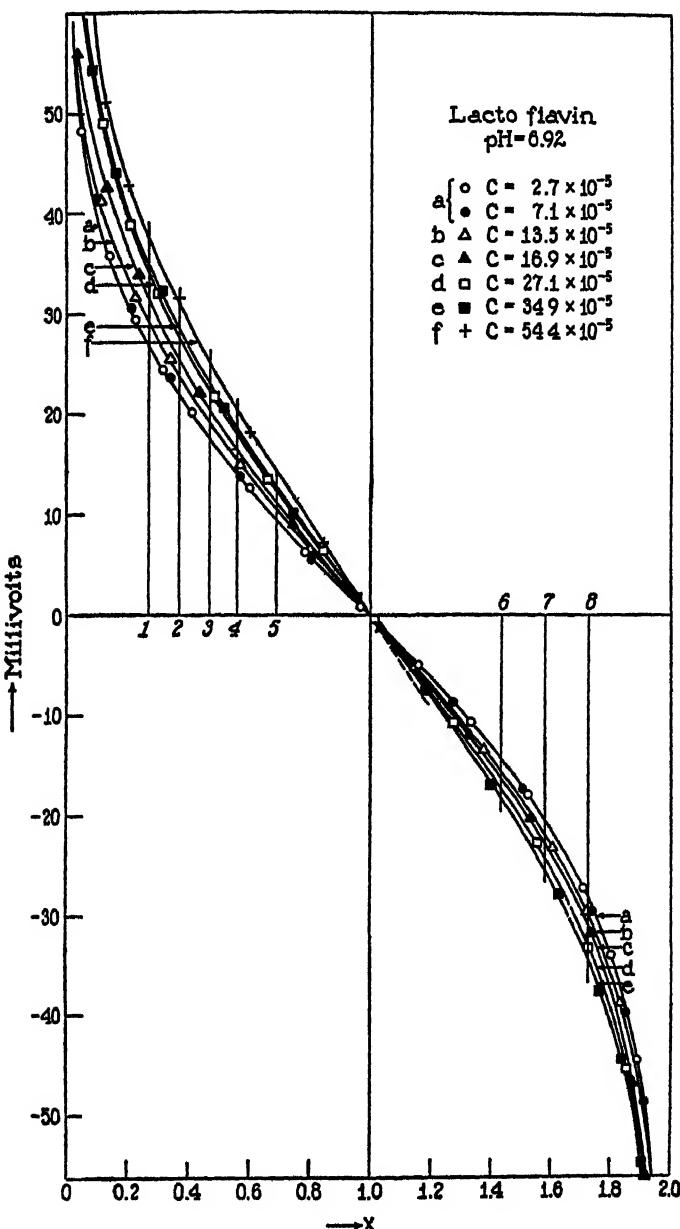


FIGURE 15. Titration curves for lactoflavin (riboflavine) for different initial concentrations of the dye.

X resemble a quinonoid structure. This is why these radicals deserve the name "semiquinones." There are several more pairs of equivalent structures, by placing the odd electron at some other C atom than in IX or X; they are of less importance and may be neglected. So, this radical, in an acid solution, has every chance of being a rather stable compound. But in an alkaline solution (formula V) this chance is lost, and in fact, the radical cannot be shown to exist at all.

The stability of the radicals of the type VII, called Wurster's dyes, is even increased by methylating the amino groups. This is the more

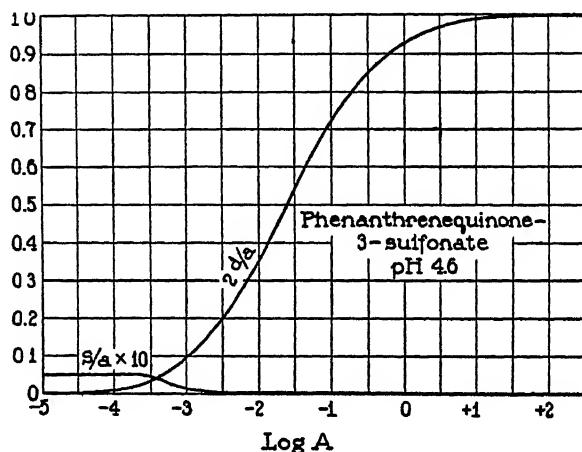
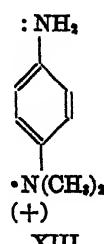
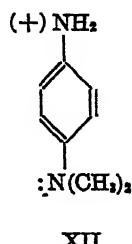
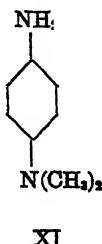


FIGURE 16. The maximum ratio of semiquinone to total dye, (s/a), and the maximum ratio at which the dye can exist in the dimeric intermediate form ($2d/a$), plotted against the log of the total concentration of the dye. Conditions the same as in FIGURE 13.

remarkable as the stability of the corresponding diimines is diminished by methylation, in contrast.

It is important to insert at this point what is meant by a pair of equivalent structures. Let us take such a diamine with partially and unsymmetrically substituted amino groups, such as XI. Its radical, in acid solution, would be XII in resonance with XIII. XII and XIII



are not entirely equivalent. Why should the resonance be so efficient in stabilizing this radical? Now, the exchange of electrons resulting in resonance stretches only from one N across the conjugated double bonds of the ring to the other N atom, but not outside this region. So, it is of little importance whether the atoms attached to the two N atoms are all alike or not, *provided there are two atoms attached at all to each N atom*, in addition to the attachment at the ring. The claim for equivalence in structure for two limiting structures between which resonance takes place, is restricted to the spacial region of the resonance. In fact, the radical XII is a very stable one.

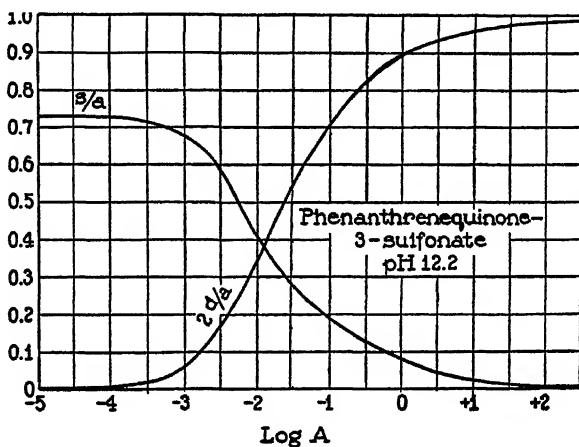


FIGURE 17. Similarly as in FIGURE 16. Conditions the same as in FIGURE 14.

Those electrons which are exchanged in this type of resonance are all in a state in which their wave-function is not spherical-symmetric; they are π electrons. For this reason, resonance can take place only if there is one definite spacial arrangement: the ring, the two N atoms, and all atoms attached to the two N atoms, must lie in one plane. Even a small deviation from a coplanar structure is liable to decrease the stability to a great extent. In the types of semiquinone radicals just now described, there is no hindrance for such a coplanar arrangement. In certain cases, where there is such a hindrance, the stability of the radical is, in fact, enormously decreased. Two examples may be shown. If in the radical XII, one H atom of the ring is substituted by a methyl group in ortho position to NH_2 , it scarcely influences the stability. If the CH_3 group, however, is ortho to the $\text{N}(\text{CH}_3)_2$ group, the radical is very unstable. The voluminous methyl group displaces

the $\text{N}(\text{CH}_3)_2$ group, also being very voluminous, from its coplanar location. Furthermore, when all four H atoms in the ring are substituted by CH_3 , but the two NH_2 groups are unsubstituted, a stable radical can be obtained. When, however, some or all H atoms of the NH_2 groups are also substituted by CH_3 , no radical exists at all, due to steric hindrance with respect to the possibility of a coplanar arrangement.

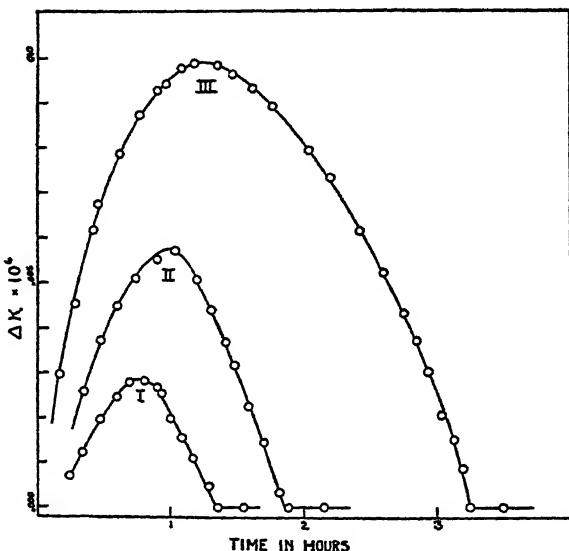


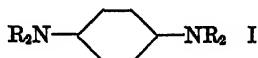
FIGURE 18. Change of magnetic susceptibility in time, for an alkaline solution of phenantrene-quinone-3-sulfonate, after addition of glucose. Three experiments with different concentrations of the dye.

CLASSIFICATION OF SEMIQUINONES

So many semiquinone radicals have been found that it would not be feasible to enumerate them all. Only the various classes will be briefly mentioned. In the following list the structure of the parent substance from which the radical is derived by a univalent step of either oxidation or, in other cases, reduction, will be given. The one of the two parent substances will be mentioned which is easiest to obtain as durable chemical preparation.

Semiquinones can be obtained by:

- 1) Oxidation of aromatic diamines in slightly acid solution (I). The four R's may be different from each other, and may be H, or alkyl, or phenyl groups, or even CH_3COOH .



2) Reduction of suitable paraquinones in alkaline solution. The best example is duroquinone (II). (In benzoquinone, secondary reactions occur in alkaline solution giving rise to complicated secondary reactions.) The whole class of anthraquinone compounds may be included in this group (II-a).

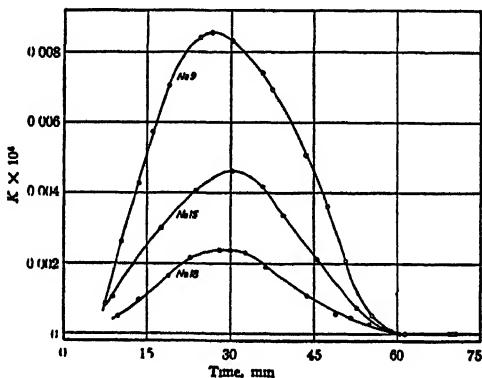
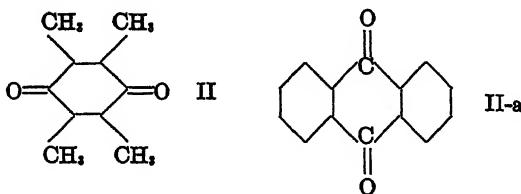
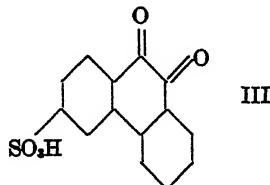


FIGURE 19. Change of magnetic susceptibility in time, for an alkaline solution of duroquinone, three experiments with varied concentration of the dye, and concentration of glucose adapted so as to stretch the period of reduction to 60 minutes approximately.

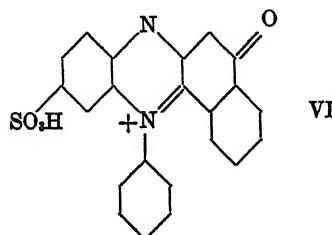
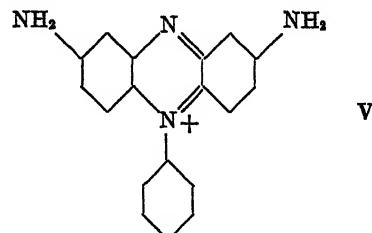
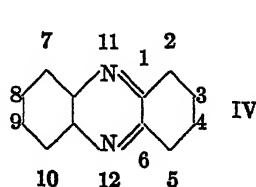


3) Reduction of suitable orthoquinones, such as phenanthrenequinone sulfonate (III) and α -naphthoquinone sulfonate.

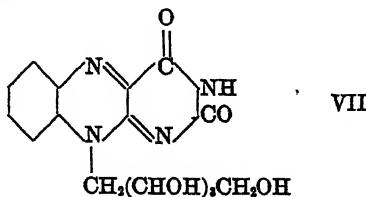


4) Reduction of phenazine (IV) in acid solution. Also many derivatives of phenazines, e. g. 2-oxy-phenazine; 2 oxy-12-methyl-

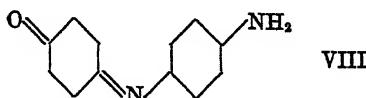
phenazinium (pyocyanin); 3, 8-diamino-9-methyl-phenazine (neutral red, in very acid solution); 3, 8-diamino-12-phenylphenazinium (safranine (V) in very acid solution) and many others; also rosindulin GG (VI) is related to this group.



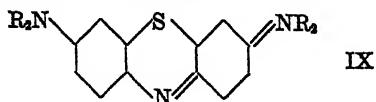
5) Reduction of flavin dyestuffs, especially also lactoflavin (riboflavin, vitamin B₂) (VII).



6) Reduction of indamines and indophenols of suitable structure, *e. g.* phenol blue (VIII).



7) Reduction of thiazines, such as thionine and methylene blue (IX), in very strongly acid solution.



8) Reduction of benzoine (X) in alkaline solution.

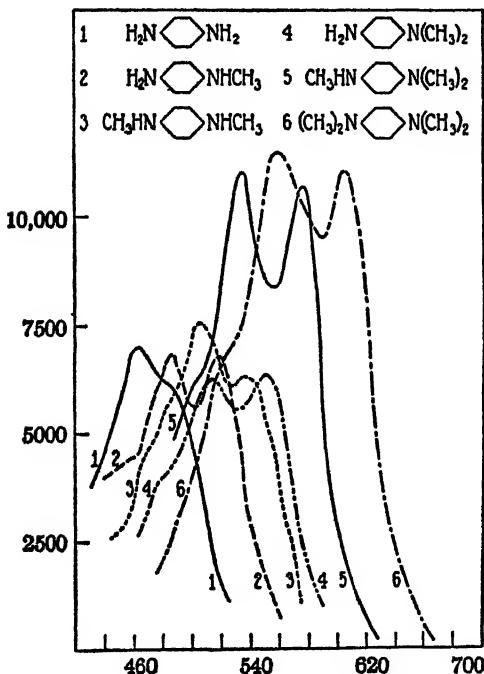
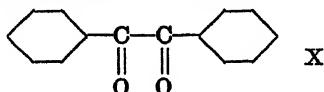
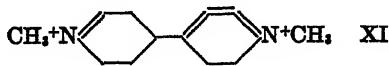


FIGURE 20. Absorption spectra of various Wurster's dyes, obtained by univalent oxidation of the aromatic diamines as indicated, all in slightly acid solution. Abcissa: wave length in $\mu\mu$. Ordinates: molar extinction coefficient. Notice the bathochromic effect of substitution in the amino group.

9) Reduction of N, N' dialkyl γ dipyridilium salts (viologens).

THE PROBLEM OF DIMERIZATION

The occurrence of dimerization, or the dissociation of a large molecule consisting of two asymmetrical moieties, into two radicals, has been known for a longer time than the dismutation. The first known case was Gomberg's discovery that hexaphenylethane can dissociate

into two electroneutral triphenylmethyl radicals, although only in quite water-free, non-polar, organic solvents. The dismutation of this radical is more difficult to accomplish than in the water-soluble semiquinone radicals.

It is quite natural that the dimerization of two radicals so as to form a valence saturated, diamagnetic compound is rather general. However, the constant of this equilibrium varies very widely. Since

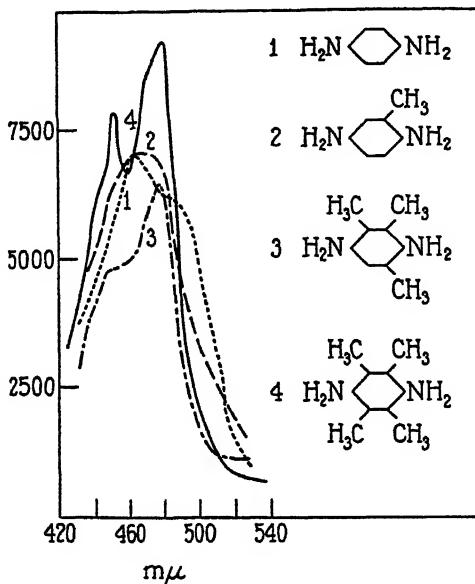
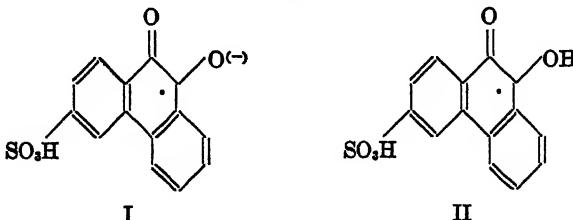


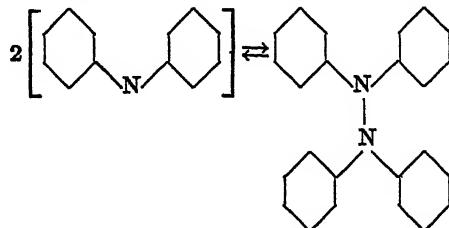
FIGURE 21. Another series of absorption spectra of Wurster's dyes, showing that substitutions at the ring have no bathochromic effect but a quite different influence on the pattern of the bands.

dimerization is a bimolecular reaction, its degree depends largely on the concentration. A great many of the radicals have so slight a tendency for dimerization that in ordinary concentrations, such as are used in potentiometrical or optical methods, dimerization is entirely unnoticeable. It is especially interesting that dimerization is much smaller in such a state of ionization when equivalent resonance prevails, than in a state where resonance is non-equivalent. So, for instance, in phenanthrenequinone sulfonate, in alkaline solution, where the radical has the structure I with equivalent resonance, dimerization begins to become noticeable only in higher concentration, whereas in acid solution, where there is no equivalent resonance (II),

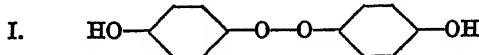
in low concentration no intermediate form to any safely measurable extent is formed, whereas in higher concentration the intermediate form does appear, yet entirely as a dimeric compound.



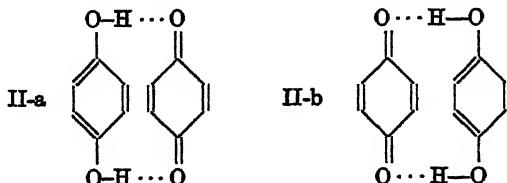
Concerning the structure of the dimeric forms, it may vary according to constitution. When there is a "bivalent N atom," the dimeric will be often of the hydrazine type



However, when there is a univalent oxygen atom, the analogous peroxide structure is probably never formed. Here something quite different happens. The radical intermediate between quinone and hydroquinone dimerizes, although only in the solid state, to the well known quinhydrone, an intensely colored, diamagnetic substance. It has certainly not the peroxide structure I, and its real structure



has been very much discussed. There is, however, a good reason to assume a structure resembling very much that proposed by Willstaetter many years ago, which in more modern language may be interpreted as follows. There is one molecule of hydroquinone and



one molecule of quinone (II-a), linked by what may seem to be a two-fold hydrogen bond. It is, however, not a regular hydrogen bond. An ordinary hydrogen bond may be considered as a shared proton. Here, however, a full H atom is shared (rather two H atoms are shared). That is to say, the structure IIa is in resonance with another, IIb, in which the hydroquinone and the quinone trade places. In order that this bond may be formed, compatible with permissible valence angles and atomic distances, the two rings must lie flat upon each other. Evidence for this assumption is the fact that duroquinone + durohydroquinone (different from the others only by four CH₃ groups substituting the four H atoms at the ring) do not form any quinhydrone; they cannot lie close enough upon each other, due to steric hindrance.⁴

The Role of the Semiquinones in the Kinetics of Oxidation-Reduction

During the last years, in all classes of the more familiar reversible oxidizable and reducible organic compounds, all being dyestuffs or at least colored substances, the formation of a semiquinone radical, even in aqueous solution, to a well measurable amount has been demonstrated. Quite recently we succeeded in showing the existence of the radical also for thiazines, such as methylene blue, which presented special technical difficulties. Now we may justly assert that radicals have been shown to exist in all the more familiar classes of reversible bivalent oxidation-reduction systems. In irreversible systems, such as alcohol-aldehyde, no intermediate radical has been demonstrated to exist in equilibrium with its parent substances, at least not to an analyzable amount. We can conclude herefrom, that the semiquinone formation constant must not be too small if the system should behave as a reversible one. Any bivalent oxidation must and can proceed only in univalent steps. Provided the radical formation constant is not too small, it does not matter how large it is; it may be 1000 or 0.01; the concentration of the radical during oxidation-reduction in this case is not the limiting factor for the rate of reaction. If, however, this constant becomes too small, the concentration of the radical may become the limiting factor of the rate, and then the process will be sluggish. In this case, the energy of formation of the intermediate radical is so high that its formation represents the essential part of the activation energy of the bivalent oxidation. When, in neutral or acid solution, alcohol is to be oxidized, the potential range of the oxidizing agent must be positive enough to form the radical; and if aldehyde is

⁴ Michaelis, L., Schubert, M. P., Reber, E. K., Kuck, J. A., & Granick, S. Jour. Am. Chem. Soc. 60: 1678 1938.

to be reduced, the potential range of the reducing agent must be negative enough to form the radical. The difference of the potential range of that oxidizing agent just strong enough to oxidize alcohol at a measurable speed, and that of the reducing agent just strong enough to reduce aldehyde with a measurable speed, is the difference between what Conant and Fieser have called the apparent oxidation potential of alcohol, and the apparent reduction potential of aldehyde. In a reversible system, both coincide, in other words, there is no over-voltage either in reduction or in oxidation. The role of a catalyst, or a respiratory enzyme, is to diminish the energy necessary to form the intermediate radical. For this purpose, the catalyst should be able to form a compound with the substrate in which the equilibrium between the oxidized form of the substrate, the reduced form, and the intermediate radical, is more in favor of the radical than in the uncombined substrate itself. Although this idea still needs further experimental support, it has in any case the advantage of reducing a problem of kinetics to one of thermodynamics.

REFERENCES

Reviews

Michaelis, L.

1935. Chem. Rev. 16: 243-286.
 1937. Trans. Electrochem. Soc. 71: 107-125.

—; & Schubert, M. P.

1938. Chem. Rev. 22: 437-470.
 [Theory of two-step oxidation.]

—; & Smythe, C. V.

1938. Ann. Rev. Biochem. 7: 1-36.

Weitz, E.

1928. Z. Elektroch. 34: 538.

Theory of Two-step Oxidation

Elema, B.

1933. Jour. Biol. Chem. 100: 149.
 1935. Rev. trav. chim. 54: 76.

Geake, A.

1938. Shirley Inst. Mem. 16: 111.

Michaelis, L.

1932. Jour. Biol. Chem. 96: 703.
 [Comprehensive theory.]
 —; & Schubert, M. P.
 1937. Jour. Biol. Chem. 119: 133.
 [Points of inflection.]

Magnetism

- Katz, H.**
 1933. Z. Physik. 87: 238.
- Klemm, W.**
 1936. Magnetochemie. Leipzig.
- Kuhn, R.**
 1939. Z. angew. Chemie 46: 478.
 [Paramagnetism in solid state.]
- ; & Ströbele, R.
 1937. Ber. 70: 753.
 [Magnetism of flavins in solid state.]
- Michaelis, L., Boeker, G. F.; & Reber, R. K.**
 1938. Jour. Am. Chem. Soc. 60: 202.
 [Paramagnetism in solution.]
- , Reber, R. K.; & Kuck, J. A.
 1938. Jour. Am. Chem. Soc. 60: 214.
 [Paramagnetism in solution.]
- , Schubert, M. P., Reber, R. K., Kuck, J. A.; & Granick, S.
 1938. Jour. Am. Chem. Soc. 60: 1678.
 [Paramagnetism in solution.]
- Rumpf, P.; & Trombe, F.**
 1938. Jour. Chim. Physique 35: 110.
 [Paramagnetism of Wurster's dyes.]

Special Chemistry of Semiquinone Radicals

- Clemo, G. R.; & McIlwain, H.**
 1934. Jour. Chem. Soc. 1991.
 [Phenazin.]
- Elema, B.**
 1933. Rec. trav. chim. 52: 569.
 [Chlororaphin.]
- ; & Sanders, A. C.
 1931. Rec. trav. chim. 50: 807.
 [Pyocyanin.]
- Friedheim, E. A. H.**
 1931. Jour. Exp. Med. 54: 207.
 [Pyocyanin.]
1933. Biochem. Z. 259: 257.
 [Hallachrom.]
- ; & Michaelis, L.
 1931. Jour. Biol. Chem. 91: 355.
 [Pyocyanin.]

Geake, A.; & Lemon, J. T.

1938. Shirley Inst. Mem. **16**: 111 and 125.
[Anthraquinones.]

Hantzsch, A. W.

1916. Ber. **49**: 511.
[Phenazonium salts.]

—; & Burawoy, A.

1932. Ber. **65**: 1059.

Hill, F. S.

1936. Proc. Soc. Exp. Biol. Med. **35**: 363.
[Phthiocol.]

Kögl, F.; & Tönnis, B.

1932. Ann. **497**: 265.
[Chlororaphin.]

Kuhn, R.; & Rudy, H.

1934. Ber. **67**: 1298.
[Flavins, riboflavin.]

Michaelis, L.

1931. Journ. Am. Chem. Soc. **53**: 2953.
[Wurster's dyes.]
1931. Jour. Biol. Chem. **92**: 211.
[Pyocyanin, oxyphenazin, rosinduline.]
1936. Jour. Am. Chem. Soc. **58**: 873.
[Beta-naphtoquinone-sulfonate.]
1936. Jour. Am. Chem. Soc. **58**: 1916.
[Neutral and safranine.]

—; & Fetcher, E. S.

1937. Jour. Am. Chem. Soc. **59**: 1246.
[Benzoin.]
1937. Jour. Am. Chem. Soc. **59**: 2460.
[Radical and Dimer of phenanthrene-quinone-sulfonate.]

—; & Hill, E. S.

1933. Jour. Am. Chem. Soc. **55**: 1481.
[Wurster's dyes, Phenazine, Viologens.]
1933. Jour. Gen. Physiol. **16**: 859.
[Viologens.]

—, —; & Schubert, M. P.

1932. Biochem. Z. **255**: 65.
[Pyocyanin and oxyphenazin.]

—; & Schubert, M. P.

1937. Jour. Bio. Chem. **119**: 133.
[Phenanthrerequionce sulfonate.]

- , —; & Smythe, C. V.
1936. Jour. Biol. Chem. 116: 587.
[Flavins.]
- , —; & Granick, S.
1939. Jour. Am. Chem. Soc. 61: 1987.
[Wurster's salts.]
1939. Science 90: 422; Jour. Am. Chem. Soc. 62: 204.
[Thiazines.]
- , —, Reber, R. K., Kuck, J. A.; & Granick, S.
1938. Jour. Am. Chem. Soc. 60: 1678.
[Duroquinone.]
- ; & Schwarzenbach, G.
1938. Jour. Biol. Chem. 123: 527.
[Riboflavin.]
- Piccard, J.
1911. Ann. 381: 351.
1913. Ber. 46: 1843.
1926. Ber. 59: 1438.
[Wurster's dyes.]
- Preissler, P. W.; & Hempeiman, L. H.
1937. Jour. Am. Chem. Soc. 59: 141.
[Phenazine derivatives.]
- Schwarzenbach, G.; & Michaelis, L.
1938. Jour. Am. Chem. Soc. 60: 1667.
[Indophenols.]
- Stern, K. G.
1934. Biochem. Jour. 28: 949.
[Flavin.]
- ; & Holiday, E. R.
1934. Ber. 67: 1104, 1442.
[Flavins.]
- Willstätter, R.; & Piccard, J.
1908. Ber. 41: 1458.
[Wurster's dyes.]

QUANTUM MECHANICAL BASIS OF THE STABILITY OF FREE RADICALS

By G. W. WHELAND

From the Department of Chemistry, University of Chicago, Chicago, Illinois

Ever since the discovery of triphenylmethyl by Gomberg in 1900,¹ chemists have tried to find a satisfactory explanation of why, in this and a few other similar cases, a carbon atom should be content to remain trivalent, when it could be tetravalent just as well as not. Of the various explanations that have been proposed, there are only two that I wish now to say anything about. The first of these involves the idea of steric hindrance. The phenyl groups are comparatively large, and it is entirely possible that two triphenylmethyl radicals are prevented by their geometrical requirements from coming close enough together to form a strong bond. What little experimental evidence there is suggests that such an explanation is probably correct to a certain extent, but that it is insufficient to account for more than a part of the stability of the radical. For example, Bent² has estimated, from the heat of hydrogenation of hexaphenylethane to triphenylmethane, that the central carbon-carbon bond in the former has been weakened by these steric factors, as compared with the corresponding bond in ethane, to the extent of perhaps 30 kg. cal. per mole. This is an appreciable effect, but it is only a fraction of the total strength of the carbon-carbon bond in ethane, which is about 70–100 kg. cal. per mole. Apparently, then, there must be some additional factor involved, which is largely responsible for the relative stability of the triphenylmethyl radical. The same conclusion is suggested by the further fact that hexa-*p*-biphenylethane is much more highly dissociated than hexaphenylethane, although it is difficult to see how the steric effects of the *p*-biphenyl and the phenyl groups could be very different.

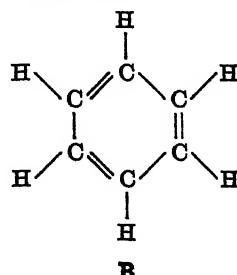
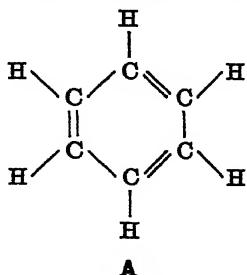
The second explanation that has been offered for the stability of free radicals is the one which I especially wish to discuss here. It is based upon a concept which has been attracting more and more attention during the last ten years and is now coming to be recognized as an important addition to the classical structural theory. I refer, of course, to the idea which is known in this country chiefly as resonance, and in England chiefly as mesomerism. I shall use only the word

¹ Gomberg, M. Jour. Am. Chem. Soc. 22: 757. 1900.

² Bent, H. E., & Cuthbertson, G. R. Jour. Am. Chem. Soc. 58: 170. 1936.

resonance, although I do not wish to imply thereby that mesomerism is not just as satisfactory a name for the phenomenon. Strictly speaking, resonance is a quantum mechanical effect, and a complete understanding of it implies some knowledge of quantum mechanics. This need not cause alarm, however, to those who lack either the time or the inclination to master that particular branch of mathematical physics. After all, the valence bond itself is also, in the last analysis, a quantum mechanical effect, but most of its important properties were well known and understood long before quantum mechanics had ever been heard of. Just as in the case of the valence bond, resonance can be talked about in non-mathematical language, and its qualitative features are not difficult to understand. As a matter of fact, much of the theory of resonance had already been anticipated by the organic chemists several years before its quantum mechanical basis was recognized.³

The modification that resonance introduces into the classical structural theory can be expressed quite simply. Whenever, for a given molecule, it is possible to write two or more structures which correspond to the same relative positions of all the nuclei but which differ in the disposition of the valence electrons, then the true structure cannot be any of these but must be of an intermediate or hybrid type. The molecule is then said to resonate among the various structures involved. Since there is frequently some misconception as to what is meant by the statement that a molecule is intermediate between, say, two structures, I wish now to digress for a short time and to discuss a specific example in some detail. I will take the case of benzene, since it is especially simple and since it is of importance for what follows. This molecule resonates principally between the two Kekulé structures, *A* and *B*, so that its structure is, in a sense,



³ For review articles by the three leading proponents, see Arndt, F. Ber. 63: 2963. 1930. Ingold, C. K. Jour. Chem. Soc. 1933: 1120; Chem. Rev. 15: 225. 1934. Robinson, R. "Outline of an Electrochemical (Electronic) Theory of the Course of Organic Reactions." Inst. Chem. Great Britain and Ireland, London. 1932.

intermediate between *A* and *B*. That does not mean that some of the molecules have structure *A* and the rest have structure *B*, nor does it mean that any given molecule spends part of its time in structure *A* and the rest in structure *B*. If such were the case, there would then be no difference between resonance and tautomerism. The correct statement is that all the molecules have the same structure, and that any given molecule keeps this one structure all the time, but that this structure is intermediate between *A* and *B*. At first sight, it seems impossible to visualize a structure of this sort. Actually it is impossible to represent such a structure by use of the conventional symbols. However, it must be remembered that these symbols are purely conventional, and that they are in no sense pictorial representations of what the molecules in question really look like. Instead they are merely short-hand abbreviations which correspond to more or less definite motions of the electrons and of the nuclei and to more or less definite average distributions of electric charge. When we say, then, that benzene has a structure intermediate between *A* and *B*, we mean that the actual motions of the various particles and the actual average distribution of charge is intermediate between the motions and distributions that correspond to the structures *A* and *B*.

So much for what resonance is. Let us now consider what its effect is upon the properties of the molecules in which it occurs. There are quite a number of these effects, which involve both the physical and the chemical properties, but there is only one that I want to say much about. This is the effect upon the stability—that is, upon the energy-content—of the molecule. It follows from the basic principles of quantum mechanics, and it has been verified experimentally in numerous cases,⁴ that, when a molecule resonates among two or more structures, it is more stable than any one of those structures alone. For example, to return to the case of benzene, we can predict the energy content of a Kekulé structure with the use of the so-called bond energies. It has been found empirically that a bond of a given type makes a fairly constant contribution to the energy content of any molecule in which it occurs. There are, of course, deviations from this rule, but they are small compared with the resonance effects that we are now discussing. The energy of a Kekulé structure is then the sum of three times the energy of a carbon-carbon single bond, plus three times the energy of a carbon-carbon double bond, plus six times the energy of a carbon-hydrogen bond.

⁴ A survey is given by Pauling, L. "Nature of the Chemical Bond". Cornell Univ. Press, Ithaca, N. Y. Chap. 4. 1939.

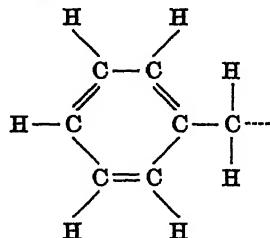
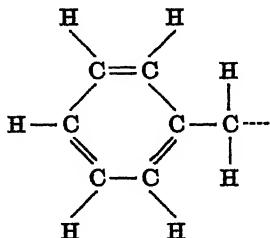
The actual energy of benzene, however, which can be determined from its heat of combustion in conjunction with other thermochemical data, is lower than this quantity by some 35–40 kg. cal. per mole. A more accurate estimate of the resonance effect is probably to be gained from a consideration of the heats of hydrogenation.⁵ It has been found that the heat of hydrogenation of an olefine is practically constant, provided that the compounds to be compared with each other have the same degree of substitution in the neighborhood of the double bond. Moreover, it has also been found that the heat of hydrogenation of a substance with more than one double bond is equal to the sum of the heats characteristic of the individual bonds, provided that these can be considered not to interact with each other through resonance. For a Kekulé structure of benzene, then, the heat of hydrogenation would be just three times that of cyclohexene. The actual heat of hydrogenation of benzene, however, is lower than this figure by 36 kg. cal. per mole. This difference between the observed energy content and that anticipated for one of the resonating structures is called the resonance energy. As is clear from this example, the resonance energy can be of quite appreciable magnitude, although in other cases it may be comparatively small.

We are now in position to consider in a qualitative way the explanation of the stability of free radicals which is presented by the theory of resonance.⁶ It will be convenient to consider first the especially simple case of dibenzyl. Actually this substance shows no tendency under ordinary conditions to dissociate into two benzyl radicals, but the same reasoning that applies in the case of hexaphenylethane suggests that here too the central carbon-carbon bond should be materially weakened, in comparison with the corresponding bond in ethane. It is not known definitely how much energy is required to dissociate ethane into two methyl radicals, but a value of somewhere between 70 and 100 kg. cal. per mole seems to be indicated by various sorts of data. As a first approximation we might consider that the same amount of energy would be required to dissociate dibenzyl into two benzyl radicals. However, in this case, the energy contents both of the undissociated molecule and of the resulting free radicals are modified by the occurrence of resonance, and so there are further factors to be considered that did not arise with ethane. In dibenzyl, each of the phenyl groups resonates between the two Kekulé structures, and the total resonance energy is just twice that of benzene,

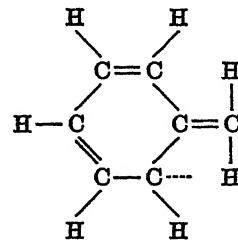
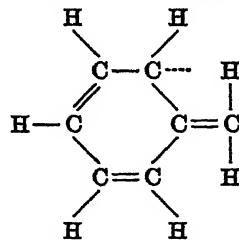
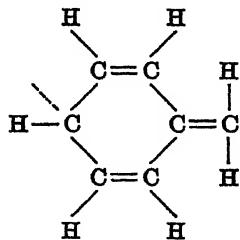
⁵ Conant, J. B., & Kistiakowsky, G. B. *Chem. Rev.* 20: 181. 1937.

⁶ Pauling, L., & Wheland, G. W. *Jour. Chem. Phys.* 1: 362. 1933.

or some 70–80 kg. cal. per mole. Each of the free benzyl radicals also resonates between the two Kekulé structures, *C* and *D*, so that, if this were the whole story, the total resonance energy of the products of dissociation would be just the same as that of the undissociated



dibenzyl. Resonance would then have no effect upon the extent of dissociation. A benzyl radical, however, has several further possibilities for resonance, which do not occur in dibenzyl. Thus, resonance in the radical involves not only the two Kekulé structures, in which the free valence is upon the methyl carbon atom, but also three additional structures, *E*, *F*, and *G*, in which the free valence is, respectively, upon

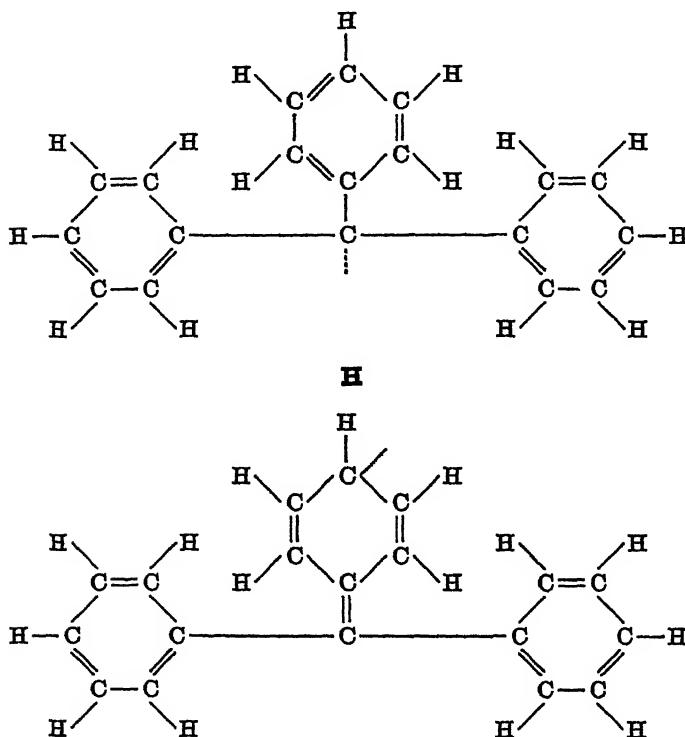
**E****F****G**

the one para and the two ortho positions. Resonance with these further structures increases the resonance energy, so that the total resonance energy of a benzyl radical is greater than that of benzene.

From this reasoning, we conclude that the dissociation of dibenzyl into free radicals is accompanied by an increase in resonance energy. Since the effect of such resonance energy is always to stabilize the system in which it occurs, we conclude further that the resonance promotes the dissociation by stabilizing the radicals to a greater extent than the undissociated molecule.

The situation is quite similar with hexaphenylethane. The undissociated molecule in this case has six phenyl groups, so that its reso-

nance energy is just six times that of benzene. Each triphenylmethyl radical has three phenyl groups, so that its resonance energy would be just three times that of benzene, if resonance were restricted to the Kekulé structures of the type *H* in which the free valence is on the



methyl carbon atom. Under such circumstances, the resonance would, as before, have no effect upon the extent of dissociation. Actually, however, resonance cannot be restricted to these structures but occurs also with quite a number of others of the type *I*, in which the free valence is upon one of the para or ortho carbon atoms of one of the rings. In the same way as with dibenzyl, it can be shown that the effect of resonance with these additional structures is again to promote dissociation into free radicals. Moreover, it can be shown that this effect is greater than with dibenzyl. If the resonance were restricted to those structures in which the free valence is either upon the methyl carbon atom or upon one of the para or ortho carbon

atoms of some particular benzene ring, the situation would be essentially the same as with dibenzyl and the extent of dissociation would also be the same (except for the effect of the steric factors that were mentioned at the outset). Since no such restriction is possible, however, resonance occurs also with structures in which the free valence is in one of the two remaining rings, and this additional resonance stabilizes the radical still further.

The foregoing qualitative discussion makes it clear that, from a consideration of the possibilities for resonance, we could expect the extent of dissociation of an ethane to increase with the number of phenyl substituents. This is, of course, in agreement with the experimental facts, but before we can be sure that resonance provides a satisfactory explanation of these facts, we need to assure ourselves that its effect is not only in the right direction but also of the right order of magnitude. The central carbon-carbon bond has been weakened in passing from ethane to hexaphenylethane to such an extent that the energy required to break it is decreased from about 70–100 kg. cal. per mole to only about 10 kg. cal. per mole. Of this decrease of 60–90 kg. cal. per mole, the steric factor apparently accounts for about 30, so that resonance must account for a further 30–60 kg. cal. per mole if it is to be a satisfactory explanation of the facts. Fortunately, it has been possible in several cases to carry through an approximately quantitative calculation of the resonance energies of the undissociated molecules and of the free radicals, and the results have been in satisfactory agreement with experiment. We can, accordingly, feel fairly confident that we now understand the two most important of the factors that are responsible for the stability of the free radicals of the triphenylmethyl type. There are doubtless a number of small effects which cause small differences in stability and which are still unrecognized, but steric hindrance and resonance, between them, account for practically all of the observed weakening of the bond in hexaphenylethane.

In the space that remains, I want to say something about the more quantitative methods that have been used in estimating the effect of resonance upon the stability of free radicals. Anything approaching a rigorous solution of the quantum mechanical problem is entirely out of the question on account of its extraordinary mathematical complexity. Consequently, it is necessary to introduce a number of approximations and simplifying assumptions if any progress at all is to be made. There are two different methods that have been used for this purpose, and these differ considerably in the nature of the

approximations and assumptions upon which they are based. However, they also have certain rather fundamental similarities and, as a matter of fact, they lead to nearly identical answers to most of the problems to which they have been applied.

One of these methods, the so-called molecular orbital method,⁷ I do not intend to say much about. My reason for neglecting it is not that it is uninteresting or unimportant, nor even that its mathematical complexities are too difficult to be discussed here. As a matter of fact, it is just as interesting and just as important as the second method, and its mathematical development is considerably simpler. However, it does not make explicit use of the concept of a valence-bond structure, and, consequently, its relationship to the qualitative discussion that I have just given is not at all apparent.

The second of the two approximate methods, the so-called valence-bond method,⁸ is the one that I wish to discuss. As its name implies, it does make explicit use of the structural concept, and, for that reason, it can be somewhat more easily grasped by those who have a chemical, especially an organic chemical background. I shall have to commence by making a few general remarks about the basic principles of quantum mechanics. The quantum mechanical description of any atom, molecule, or other similar system is contained in what is called the eigenfunction of the system. This eigenfunction is a more or less ordinary mathematical function of the spatial coordinates of all the particles of which the system is composed. There are also parameters that refer to the spins of the particles as well, but we need not worry about these here. The eigenfunction, for example, that describes a hydrogen atom in its ground state is simply e^{-ar} , where e is the logarithmic base, 2.718...; a is a constant, given in terms of the natural mathematical and physical constants like π , h , and so on; and r is the distance between the proton and the electron. In the general case, the eigenfunctions are much more complicated, and, for even such a simple molecule as H_2 , several pages would be required to write out in detail the best available approximation to the function. The eigenfunctions are obtainable, in principle, by solving a certain partial differential equation known as the Schrödinger equation, or the wave equation, but this can be done, in practice, in only a few especially simple cases. Usually, the best that one can do is to obtain an approximate eigenfunction by the use of approximate methods of solution. Before discussing these approximate methods, I want to say something about

⁷ For the application of this method to the question of dissociation into free radicals, see Huckel, E. Zeits. f. Physik. 83: 632. 1933.

the significance of the eigenfunction, that is, about the way in which it is related to the state of the atom or molecule that it describes. Without going into all the details, I shall mention only the fact that, if one knows the eigenfunction for any system, he can then calculate by straightforward, although often very complicated mathematical methods, the numerical value (or, at any rate, the average numerical value) of any property of the system that can be measured experimentally. In particular, from a knowledge of the eigenfunction of a molecule, one can calculate its energy. Of more interest to us is the fact that from a knowledge of an approximate eigenfunction, one can calculate an approximate value of its energy. The closer the approximate eigenfunction is to the correct one, the closer, of course, the approximate energy is to the correct one.

Now let us consider the method for setting up an approximate eigenfunction for a molecule like benzene. There is a definite procedure, due originally to Slater,⁸ by which we can set up an approximate eigenfunction for either one of the two Kekulé structures, or, in general, for any valence-bond structure of any molecule. Such a function is not perfect, of course, but there is reason to believe that it is fairly good, or that, at any rate, it is good enough for our present purposes. I do not wish to go into the exact mathematical form of Slater's functions, because they are quite complex and are of no immediate interest to us here. However, I shall designate the approximate eigenfunctions of the Kekulé structures *A* and *B* by the symbols ψ_A and ψ_B , respectively. Neither ψ_A nor ψ_B is an accurate representation of benzene, and so it seems natural to try a more general function of the form $\Phi = a\psi_A + b\psi_B$, where *a* and *b* are numerical constants. We can now choose the values of *a* and *b* in such a way that Φ becomes the best possible eigenfunction for benzene that can be expressed in the form under consideration. Whenever an approximate eigenfunction is obtained in this way by making a linear combination of approximate eigenfunctions characteristic of two or more valence-bond structures, the molecule is then said to resonate among the structures involved. This is accordingly, the quantum mechanical basis of the theory of resonance, which I discussed a short time ago from a purely qualitative point of view.

The question still remains as to how the best possible values of the numerical coefficients are to be determined. It turns out to be easier to calculate directly the energy which corresponds to these best

⁸ Slater, J. C. Phys. Rev. 38: 1109. 1931.

See also, Kimball, G. E., & Eyring, H. Jour. Am. Chem. Soc. 54: 3876. 1932.

values, since that is the quantity that is of interest to us. The method for doing this is to solve an algebraic equation, known as the secular equation. With the introduction of certain further simplifications and approximations which I do not wish to discuss here, this secular equation for the case of resonance between the two Kekulé structures of benzene is found to be of the form

$$\begin{vmatrix} Q + \frac{3}{2}\alpha - W & \frac{1}{4}Q + \frac{3}{2}\alpha - \frac{1}{4}W \\ \frac{1}{4}Q + \frac{3}{2}\alpha - \frac{1}{4}W & Q + \frac{3}{2}\alpha - W \end{vmatrix} = 0.$$

Q and α are two definite integrals, defined in terms of the functions ψ_A and ψ_B , and W is the approximate energy of the molecule. On solving this for W , we find $W = Q + 2.4\alpha$ or $W = Q$. It may seem strange that there are two roots, giving two different values of the energy. The explanation is that one root gives the energy of the molecule in its ground state, and the other gives the energy in an excited state of only spectroscopic interest. In order to interpret this result, we need to know the values of the integrals Q and α . We could evaluate them numerically by actually carrying out the integrations, but that would involve a great deal of work. Moreover, there is some reason to doubt that the integration would provide very useful values, because there is considerable uncertainty in regard to the precise form of the functions that appear in the integrands. We shall, accordingly, use a more empirical procedure. If the molecule were considered not to resonate between the structures A and B but to have structure A alone, the secular determinant would consist of only the upper left-hand corner of the one above. This would give $|Q + 3/2\alpha - W| = 0$ or $W = Q + 3/2\alpha$. Similarly, if the molecule had structure B alone, the secular determinant would consist of only the lower right-hand corner and would become $|Q + 3/2\alpha - W| = 0$. Thus again, $W = Q + 3/2\alpha$, in accordance with the fact that the two Kekulé structures, being equivalent, must have the same energy. Thus we see that, while the energy of a Kekulé structure is $Q + 1.5\alpha$, resonance between the two structures changes the energy to either $Q + 2.4\alpha$ or Q . The difference, either -0.9α or 1.5α , is then equal to the resonance energy, which we have already found to be about 36 kg. cal. per mole. Thus α must be either about -40 kg. cal. per mole or else about 24 kg. cal. per mole. Since integrals of the type of α are usually found on numerical integration to be negative in sign, we shall assume that the value of -40 kg. cal. per mole is the correct one.

The difference between the normal and the excited states of benzene is thus -2.4α or 96 kg. cal. per mole. This corresponds to a wavelength of a little less than 3000 Å, which is indeed in the region of the ultraviolet absorption spectrum of benzene. (I might add that a more detailed calculation of the present type, which considers resonance with other structures than the Kekulé structures, gives a resonance energy of -1.1α instead of -0.9α . This gives $\alpha = -33$ kg. cal. per mole. To the same approximation, the first excitation energy is -2.6α or 86 kg. cal. per mole, so that the calculated position of absorption is about the same as before.)

The single result for benzene alone is of little interest unless some independent check is obtained for the value derived for the integral α . The value of -33 or -40 kg. cal. per mole is quite a reasonable one and is entirely in line with the values obtained by actual integration for similar integrals that occur in the treatment of other molecules. The fact that the calculation is in fair agreement with the spectroscopic data is also a gratifying confirmation of the value obtained for α . The most conclusive evidence that the calculation is significant and reliable, however, is provided by similar calculations that have been made for a number of further aromatic molecules. I need not go into the details of these additional calculations, since the general principles involved are the same in all cases. For naphthalene, for example, the calculated resonance energy⁹ is -2.0α . Comparison with the observed value of 75 kg. cal. per mole gives $\alpha = -37.5$ kg. cal. per mole, in satisfactory agreement with the value found to apply in the case of benzene. The extreme range in the values of α obtained from all the aromatic hydrocarbons treated so far is only from about -33 to about -40 kg. cal. per mole.¹⁰ This is a remarkably close agreement in view of the extreme crudity of the calculations.

The application of this method of calculation to the problem of the stability of the free radicals should now be apparent. We have to calculate the resonance energy of the undissociated ethane and also that of the two free radicals formed from it. The difference between these quantities is then a measure of the extent to which the dissociation is promoted by the resonance. In dibenzyl, for example, we find a total resonance energy of -1.8α , which is, of course, just twice that of benzene. (It is to be noticed that we have returned to the cruder approximation of considering resonance only among the various Kekulé structures, so that the resonance energy of benzene is only

⁹ Sherman, J. *Jour. Chem. Phys.* **2**: 488. 1934.

¹⁰ Wheland, G. W. *Jour. Chem. Phys.* **2**: 474. 1934.

— 0.9α and not — 1.1α .) To the same approximation, we find the resonance energy of one benzyl radical to be — 1.4α , so that that of the two together is — 2.8α . Thus the resonance favors the dissociation into radicals to the extent of — 1.0α , or about 35–40 kg. cal. per mole. Since there is no reason to expect the steric factor to be anything like as important here as with hexaphenylethane, this effect is much too small to permit an easy rupture of the central carbon-carbon bond in dibenzyl.

In hexaphenylethane, the treatment is similar. The resonance energy of the undisassociated molecule is here just six times that of benzene, or — 5.4α , while the total resonance energy of two triphenylmethyl radicals is — 7.6α . The resonance thus promotes the dissociation by — 2.2α , or some 77–88 kg. cal. per mole. This is rather larger than the 30–60 kg. cal. per mole that are required in order to account for the experimental data. The crudity of the calculation is doubtless partially responsible for the discrepancy, but a more significant factor is probably to be found in the fact that the resonance is partially inhibited by a steric effect that has been recognized by Michaelis with other kinds of free radicals.¹¹ The quinoid structures (of type I above) in which the free valence is located on an ortho or para position in one of the rings, require that the radical be entirely coplanar. Such a geometrical arrangement, however, cannot be achieved without bringing the ortho hydrogen atoms of the benzene rings impossibly close together. Consequently, since the radical is necessarily non-coplanar, some, at any rate, of the quinoid structures are made considerably less stable than they would otherwise be. As a result of this, the resonance is partially inhibited, and the resonance energy is correspondingly decreased. The calculation, which makes no allowance for such a steric effect, must, therefore, lead to too great a difference between the resonance energy of the ethane and that of the radicals, and so to too great a dissociation of the ethane. In any case, whether the proposed explanation of the discrepancy is the correct one or not, it is gratifying that the calculated resonance effect is of the right order of magnitude, and neither too large nor too small by some very large factor, as it might well have been.

A further gratifying feature of the calculation is that it accounts not only for the dissociation of hexaphenylethane but also for the relative effectiveness of the various aryl groups. The results can be presented most easily in tabular form. In the first column of the following table are listed the names of the radicals formed by dissociation

¹¹ Michaelis, L., Schubert, M. P., & Granick, S. Jour. Am. Chem. Soc. 61: 1981. 1939.

tion of the ethane; in the second are the differences in resonance energy between the ethanes and the corresponding pairs of radicals;

TABLE I^{7, 10}

Radical	Valence-Bond Method	Molecular Orbital Method
Phenylmethyl (Benzyl)	-1.02 α	-1 44 β
β -Naphthylmethyl	-1.26 α	-1 49 β
<i>p</i> -Biphenylmethyl	-1.29 α	-1 51 β
α -Naphthylmethyl	-1.50 α	-1 62 β
Fluoryl	-1 60 α	-2 69 β
Diphenylmethyl	-1 68 α	-2 60 β
Phenylfluoryl	-2.15 α	-3.67 β
Triphenylmethyl	-2 22 α	-3 59 β
β -Naphthylidiphenylmethyl	-2 34 α	
<i>p</i> -Biphenyldiphenylmethyl	-2.35 α	-3.64 β
Di- <i>p</i> -biphenylphenylmethyl	-2.47 α	-3.68 β
α -Naphthylidiphenylmethyl	-2.48 α	-3.71 β
Tri- <i>p</i> -biphenylmethyl	-2 58 α	-3 72 β

in the third are the same differences, as calculated by the molecular orbital, instead of by the valence-bond method. (The quantity β is again an integral, of a type more or less analogous to α , which comes into the treatment in a similar way. It also is negative in sign, but somewhat smaller in magnitude than α .) It will be noted first that the two methods of calculation give identical orders of increasing dissociation, except for the cases involving the fluoryl radical, which will be discussed later. A second point of interest is that the calculated effect of resonance increases with the number of aryl substituents, as does also the observed extent of dissociation of the ethane. And finally, the calculations reproduce the empirical order: α -naphthyl > *p*-biphenyl > β -naphthyl > phenyl, where the symbol > may be read "is more effective in promoting dissociation than." The discrepancy in the case of the fluoryl compounds is, of course, due to errors introduced into the calculations by their crudely approximate nature. It is not possible to say with much assurance which method of calculation is the more nearly correct in this case. At first sight, the valence-bond method seems better, since it gives numerical results in agreement with the observed order of dissociation, while the molecular orbital method apparently inverts the order when the fluoryl compounds are compared with the corresponding diphenyl compounds.

However, there is evidence¹² to show that the steric factor alone is more than sufficient to account for the observed difference in dissociation between hexaphenylethane and diphenyldifluoryl, so that the resonance effect, which is all that is calculated here, may after all be greater in the latter case than in the former. That such a steric difference should exist is, moreover, quite reasonable since there can be no question that a fluoryl group occupies less space than a diphenylmethyl group does.

So far, I have talked about only hydrocarbon free radicals, but the present theory is not restricted to them, at any rate, as far as its qualitative aspects are concerned. It can indeed be extended to cover radicals of practically all other types. The dissociation of tetraphenylhydrazine, for example, is analogous to that of tetraphenylethane. The reason why the dissociation proceeds to a greater extent in the former case than it does in the latter is to be found, at least partially, in the fact that less energy is required to break a single bond between two nitrogen atoms than is required to break one between two carbon atoms. The existence of compounds in which oxygen, sulfur, or some other element exhibits an anomalous valence can also be interpreted similarly in a great many cases with the use of the resonance concept. Unfortunately, however, it is impossible to carry the treatments of any of the non-hydrocarbon free radicals beyond the qualitative stage, because, up to the present time, the approximate quantitative calculations have not been extended to these more general types of radical.

The foregoing discussion should be sufficient to show the importance of the role played by resonance in connection with the stability of free radicals. There are, however, some aspects of the problem that still remain to be solved. Alkyl groups, for example, often have a marked effect upon the extent of dissociation, whether present as a substituent in an aromatic ring¹³ or attached directly to one of the ethane carbon atoms.¹⁴ It is perhaps not impossible that these may be involved in resonance in some way that is not understood at present, or that their effect is of some purely electrostatic nature. In a number of cases, they may weaken the ethane linkage by increasing the mutual steric repulsion of the substituted methyl radicals. It is evident, then, that, while we have a good picture of the broad outlines of the problem, much more work is necessary before all the details will have been filled in.

¹² Bent, H. E., & Cline, J. E. *Jour. Am. Chem. Soc.* **58**: 1624. 1936.

¹³ Roy, M. F., & Marvel, C. S. *Jour. Am. Chem. Soc.* **59**: 2622. 1937.

¹⁴ Conant, J. B. *Jour. Chem. Phys.* **1**: 427. 1933.

APPLICATION OF THE DROPPING MERCURY ELECTRODE FOR THE DETECTION OF INTERMEDIATE RADICALS*

BY OTTO H. MÜLLER

From the Department of Surgery of The New York Hospital and Cornell University Medical College, New York City

INTRODUCTION

Two-step reductions of some metallic ions have been known for a long time in polarographic work with the dropping mercury electrode. For example, the reduction of chromic ions was found to take place in two unequal steps which were separated by about 600 mv. and, therefore, easily visible.^{1, 2} The first step required one electron for the reduction of trivalent chromic to the stable, divalent chromous ion. The second step, which was twice as high, required two electrons for the reduction of the divalent chromous ion to metallic chromium. There is no polarographic evidence that this latter step can similarly be broken up into two components, and it becomes necessary to assume that, if it exists at all, the monovalent chromium ion must be extremely unstable. The polarographic curve of the reduction of chromic ions thus serves as a good example of the two extremes in stability of the intermediates in a reduction.

Of the many organic compounds which are reversibly reduced and oxidized, involving a change of two electrons, a special group may be selected in which the intermediate radicals, which have been called semiquinones, become stable enough in solutions of suitable pH to permit studies which may be either colorimetric, magnetometric, potentiometric, or polarographic. It is this group of compounds which will be discussed in this paper.

When Müller and Baumberger³ first found that reversible organic oxidation-reduction systems could be studied at the dropping mercury electrode and that their half-wave potentials in well-buffered solutions were identical with the established E_{\circ} ' values of these compounds, they concluded that it should also be possible to determine semiquinones polarographically. The only compound available to them for such an investigation was Rosinduline GG. This showed the expected

* Supported by a grant from the John and Mary B. Markle Foundation.

¹ Demassieux, M., & Heyrovsky, J. *Jour. Chim. Phys.* 26: 219. 1929.

² Prajzler, J. *Collection Czechoslov. Chem. Commun.* 3: 406. 1931.

³ Müller, O. H., & Baumberger, J. P. *Trans. Electrochem. Soc.* 71: 181. 1937.

two-step reduction,³ but the demonstration was not very convincing because of interference by a number of other polarographic waves which were, no doubt, due to impurities. Fortunately, it has been possible to continue this work through the kindness of Dr. L. Michaelis of the Rockefeller Institute for Medical Research in New York City who supplied numerous samples of other semiquinone-forming compounds of highest purity which had been prepared and studied in detail in his laboratory. For this and for many valuable suggestions, I wish to express my thanks to Dr. Michaelis.

The polarographic curves obtained in the present investigation have been free from complications and verify the conclusions reached by Müller and Baumberger.³ It has been possible to apply to the polarographic curves modifications of the equations developed by Michaelis and associates^{4, 5} in their potentiometric studies of semiquinones. Minor discrepancies exist, as will be pointed out, but, on the whole, a satisfactory agreement between polarographic and potentiometric results is found, showing the feasibility of applying the dropping mercury electrode to the detection of intermediate radicals. While the polarographic method is less accurate at present than the potentiometric method, it may nevertheless be useful on account of the speed with which preliminary studies for orientation may be carried out. The polarographic method has a definite advantage over other methods in the fact that it is still applicable in studies where the potentials fall into the over-voltage range and in studies of partially reversible systems.

THE POLAROGRAPHIC METHOD

It is impossible to give more than a short sketch of Heyrovský's polarographic method, which is relatively unknown despite the fact that it originated in 1922. For additional information see the monographs and reviews of Heyrovský,^{6, 7} Semerano,⁸ Hohn,⁹ Kolthoff and Lingane,¹⁰ and Müller.^{11, 12}

⁴ Michaelis, L. *Trans. Electrochem. Soc.* **71**: 107. 1937.

⁵ Michaelis, L., & Schubert, M. P. *Chem. Rev.* **22**: 437. 1938.

⁶ Heyrovský, J. "Physikalische Methoden der analytischen Chemie." Vol. 2. edited by W. Böttger. Akad. Verl. Ges. Leipzig. 1936.

⁷ Heyrovský, J. "Physikalische Methoden der analytischen Chemie." Vol. 3. edited by W. Böttger. Akad. Verl. Ges. Leipzig. 1939.

⁸ Semerano, G. "Il polarografo, sua teoria e applicazioni." 2nd ed. Draghi, Padova. 1938.

⁹ Hohn, H. "Anleitung für die chemische Laboratoriumspraxis." Vol. 3. edited by E. Zind. Julius Springer, Berlin. 1937.

¹⁰ Kolthoff, I. M., & Lingane, J. J. *Chem. Rev.* **24**: 1. 1939.

¹¹ Müller, O. H. *Chem. Rev.* **24**: 95. 1939.

¹² Müller, O. H. *Cold Spring Harbor Symposia. Quant. Biol.* **7**: 59. 1939.

One might say that the polarographic method is something between a potentiometric and an electrolytic method. In the *potentiometric* method, the potential of an electrode in solution is measured in a manner that avoids, as much as is possible, the flow of any current so that the composition of the solution will not be altered. Analyses are carried out by titrations, the progress of which is indicated by the electrode potentials. In the *electrolytic* method, usually the wanted component of the solution is deposited exclusively on a suitable electrode for further analysis by applying the proper electromotive force across the cell. Here a decrease in the current indicates the progress of the electrolysis. In the *polarographic* method, a potential is applied which causes a reaction at an electrode with a very small but well defined and reproducible surface. The electrolysis is never carried to completion and the solution remains essentially unaltered; the observed current is a function of the concentration of the reacting material and serves for quantitative analysis; the potential of the electrode can be calculated from the applied voltage and the *IR* drop in the circuit, and is used for qualitative analyses.

The most satisfactory electrode meeting these requirements is the dropping mercury electrode. The utmost in reproducibility is obtained because every few seconds a new electrode, exactly like the previous, grows into the solution. Each presents a fresh surface but such a small one that only a minute fraction of the electroactive material can be oxidized or reduced during its lifetime. If this electrode is placed in series with a potentiometer and galvanometer and a large unpolarizable reference electrode of known potential, the current flowing at different applied voltages can be measured. Plotting the observed currents against the applied voltages, one obtains a curve which has been called a polarogram when it is obtained automatically by means of a machine, the polarograph. If an electrode reaction occurs, "S"-shaped waves appear on these polarograms with a measurable height and a well definable point of inflection. Under suitable conditions, the former is proportional to the concentration of the reacting material and serves for quantitative determinations, while the latter is a constant, characteristic for each reacting substance, and corresponds to an applied potential which serves for qualitative determinations.

For purposes of quantitative analyses, effects due to migration and absorption are eliminated, and the reacting material reaches the electrode surface only by simple diffusion in a gradient which is created by the removal of the material at the electrode. The current obtained

when this gradient is maximal has been called diffusion current and is the only type of current which will be considered in this paper. It is limited by the quantity of material which can reach the surface of the mercury drop during its existence. Ilković¹³ has derived an equation for this diffusion current which has been verified experimentally:

$$(1) \quad I_d = 0.627 n F D^{1/2} C m^{2/3} t^{1/6}.$$

Here n is the number of electrons involved in the reduction or oxidation of one molecule of the reacting substance, F is the Faraday, D is the diffusion constant of the reacting substance, C is the concentration of the reacting substance in the body of the solution, m is the weight of mercury flowing from the capillary per second, and t is the drop time (the latter two quantities must be measured at the same potential at which the diffusion current is determined).

Of course, the diffusion current cannot be obtained until the applied voltage is sufficiently negative in the case of an electro-reduction and sufficiently positive in the case of an electro-oxidation. If that condition is not fulfilled, either no current or only a fraction of the total diffusion current will flow. When the current has reached a value which is one-half that of the diffusion current, one-half of all the reacting material at the electrode surface has reacted; if the end-product is stable it will remain at that electrode surface long enough to produce a condition in which oxidant and reductant are in equal concentrations. This condition prevails, of course, only in that very thin layer of solution which is in immediate contact with the mercury surface and which will henceforth be called the interface. The current-voltage curve has a point of inflection at this point of half-diffusion current and the corresponding potential has been called "half-wave potential."¹⁴ It is necessary to emphasize, however, that the polarographic curves are not current-potential curves, but that they are current-voltage curves. However, the applied voltage which is plotted can easily be converted to applied potential if the resistance of the system is known, by the simple formula:

$$(2) \quad E = V - IR$$

in which E is the potential of the dropping mercury electrode, V is the applied electromotive force, I is the current, and R is the resistance of the circuit. The factor IR can be neglected *only* when it approaches the limit of error of polarographic measurements which is 3-10 mv.

¹³ Ilković, D. Collection Czechoslov. Chem. Commun. 6: 498. 1934.

¹⁴ Heyrovský, J., & Ilković, D. Collection Czechoslov. Chem. Commun. 7: 198. 1935.

For instance, if R is 1000 ohms, the polarographic curve may be considered a current-potential curve only up to currents of 10^{-6} Amp.; at greater currents, IR corrections for potential become necessary for accurate work.

ANALYSIS OF CURVES WITHOUT SEMIQUINONE FORMATION

Heyrovský and Ilkovič¹⁴ were the first to analyze simple polarographic curves in detail and to develop equations for them. These analyses brought out the significance of the half-wave potential which was thereupon proposed as the most satisfactory constant for the characterization of reacting substances in polarographic work. The fact that the current is a function of the diffusion gradient of the reacting material, was expressed as follows:

$$(3) \quad I = K(C - C_0)$$

where I is the current and C the concentration in the body of the solution, while C_0 is the concentration at the interface. K is the diffusion current constant. When the diffusion current is reached, the concentration at the interface, C_0 , becomes negligibly small compared to C and

$$(4) \quad I_d = K C.$$

The concentration of the product of the reaction at the interface, C_a , is also proportional to the current I (if none existed in solution before the reaction) and Heyrovský and Ilkovič write

$$(5) \quad C_a = k I$$

where k is again a complex constant.

Now if the potentials of the dropping mercury electrode during reversible reactions follow the same laws that govern the potentials of other electrodes, the following formula must hold for all parts of the curve:

$$(6) \quad E = E_0 - (RT/nF) \ln(C_a/C_0)$$

if, for simplicity, concentrations may be used instead of activities.

Substituting in this equation from above, we find:

$$(7) \quad E = E_0 - (RT/nF) \ln(kIK/KC - I)$$

or

$$(8) \quad E = E_0 - (RT/nF) \ln(I/I_d - I) + K$$

If equation (8) holds, it is obvious that a graph of E against $\ln(I/I_d - I)$ should give a straight line with a slope equal to RT/nF . This has first been tested by Tomeš¹⁵ who found slopes of 0.056, 0.029, and 0.020 v. for mono-, di-, and tri-valent inorganic reductions at 20° C. in fair agreement with the theoretically expected values.

In the case of the half-wave potential, $E_{1/2}$, the logarithmic term in equation (8) drops out because $I = I_d - I$, so that:

$$(9) \quad E_{1/2} = E_o + K.$$

This means that the half-wave potential of the dropping mercury electrode is equal to the normal electrode potential of the reacting system plus some constant which may be positive, negative, or zero.

These same equations can be extended to cover oxidations and reductions of reversible organic systems. The electrode potential of reversible organic reactions such as



is given by the well-known equation:

$$(11) \quad E = E_o - \frac{RT}{2F} \ln \frac{[\text{Red}]}{[\text{Ox}]} + \frac{RT}{2F} \ln [K_{a_1} K_{a_2} + K_{a_1} [H^+] + [H^+]^2]$$

which is written for a two-electron change and in which K_{a_1} and K_{a_2} are the two dissociation constants of the reductant. $[\text{Red}]$ and $[\text{Ox}]$ represent the total concentration of reductant and oxidant, respectively. If the pH of the solution is kept constant by adequate buffers, this formula becomes simply

$$(12) \quad E = E'_o - (RT/2F) \ln ([\text{Red}]/[\text{Ox}]).$$

If the solution contains only oxidant at the beginning of the polarographic analysis, we can substitute from equations 3-5 and get

$$(8a) \quad E = E'_o - (RT/2F) \ln (I/I_d - I).$$

A similar equation with a plus sign before the logarithmic term will be obtained for the case when only reductant is present in the body of the solution. In both cases the logarithmic term drops out at the half-wave potential where $I = I_d - I$, so that

$$(9a) \quad E_{1/2} = E'_o + K.$$

Müller and Baumberger³ verified this conclusion experimentally and found the constant K equal to zero for simple reversible organic

¹⁵ Tomeš, J. Collection Czechoslov. Chem. Commun. 9: 12. 1937.

oxidation-reduction systems. They made a detailed study of the reactions of hydroquinone, quinone, and quinhydrone at the dropping mercury electrode. If the solution was well buffered, the half-wave potentials in all three cases were identical although the dropping mercury electrode was anode in the first case, cathode in the second case, and an indicator electrode¹⁶ in the third case. They showed further that these half-wave potentials agreed with the potentials of quinhydrone obtained potentiometrically with platinum electrodes. Thus at the continually changing surface of the mercury drops conditions must exist which are similar to those existing where thermodynamic equilibrium prevails. Although this was verified in subsequent studies,^{11, 12} it has been subjected to another test in the present investigation. Again an agreement between polarographic and potentiometric results was obtained which demonstrates the existence of thermodynamic conditions at the electrode/solution interface, produced by a flow of current, and defined by the applied potential. While the body of the solution remains essentially unchanged, electrons are transferred either from or to the electrode, changing the composition of the interface to satisfy the conditions of the applied potential. One can compare this to a titration of a solution by an oxidant or a reductant; however, in this case only the material in the interface is titrated, and it is titrated directly with electrons from the electrode. This may, therefore, be called an electron-titration of the interface.¹²

Further equations for curves in which oxidant and reductant are present in various proportions at the beginning of the electrolysis, can also be developed. However, for simplicity, our treatment of the semiquinone problem will deal solely with well-buffered solutions containing only oxidant at the beginning of the electrolysis.

ANALYSIS OF CURVES WITH SEMIQUINONE FORMATION

On the basis of the above conclusions we can develop potential equations for polarographic curves of semiquinones in close analogy to those derived by Michaelis and Schubert⁵ in their potentiometric studies.

Let us consider the following reduction



¹⁶ Müller, O. H., & Baumberger, J. P. Trans. Electrochem. Soc. 71: 169. 1937.

which takes place in two separate steps of one electron each. Here B is the oxidant (Ox), B^- the semiquinone (Sem), and B^{--} the reductant (Red). If true equilibrium exists, the potential must be given by the following equation in which [Ox] and [Red] represent the total concentration of B and B^{--} respectively.

$$(12) \quad E = E'_o - (RT/2F) \ln ([\text{Red}]/[\text{Ox}]).$$

The concentrations of Red and Ox are modified according to the equilibrium of the dismutation process



which may be expressed by the dismutation constant

$$(15) \quad x = ([\text{Red}][\text{Ox}])/[\text{Sem}]^2.$$

The reciprocal of x is called the semiquinone formation constant and is designated by k . If x is very large (k very small), the amount of semiquinone present becomes negligible and the reaction will be that of a system with no intermediate step of reduction. Our problem, however, is to arrive at an equation which defines the potential when significant amounts of semiquinones are present.

We start with a solution containing a substance in its oxidized form, which can be reduced at the dropping mercury electrode. The highest concentration of this reducible substance in the electrode/solution interface is proportional to the maximal diffusion current which can be obtained, or

$$[\text{Ox}]_{\max} = kI_d.$$

Let now a potential be applied to the electrode which will cause a reduction of this substance. Then at any point along the wave of the current-voltage curve there will be a mixture of the oxidized form (Ox), the semi-reduced form (Sem), and the totally reduced form (Red), so that

$$(16) \quad \text{Ox} + \text{Sem} + \text{Red} = kI_d.$$

A current of electrons flowing from electrode to solution changes the composition of the interface as would an added reducing agent. When the current $I = I_d$, all the oxidant has been totally reduced to reductant, but at lower values of the current, a mixture of semiquinone and reductant has been formed. Only one electron is necessary per molecule for the production of the semiquinone, while two electrons are necessary for the production of the totally reduced product. This fact can be expressed as follows

(17)
$$\text{Sem}/2 + \text{Red} = kI$$

or

$$\text{Sem} + 2 \text{Red} = 2kI$$

As Michaelis and Schubert⁵ have shown, it is possible to solve such equations for Red, Ox and Sem. Let us divide equations (15-17) by Sem; and set Ox/Sem = Ω and Red/Sem = ρ , then

(15a)
$$\Omega\rho = x$$

(16a)
$$\Omega + \rho + 1 = kI_d/\text{Sem}$$

(17a)
$$1 + 2\rho = 2kI/\text{Sem}$$

Now divide equation (17a) by equation (16a) in order to eliminate the remaining Sem:

(18)
$$\frac{1 + 2\rho}{\Omega + \rho + 1} = \frac{2I}{I_d}$$

Substituting into this equation from equation (15a) we obtain, writing only the solution with a positive sign before the square root:

(19)
$$\Omega = \frac{I}{4I} [\sqrt{(I_d - 2I)^2 + 16 I \kappa (I_d - I)} + (I_d - 2I)]$$

(20)
$$\rho = \frac{1}{4(I_d - I)} [\sqrt{(I_d - 2I)^2 + 16 I \kappa (I_d - I)} - (I_d - 2I)]$$

Since $\rho/\Omega = \text{Red/Ox}$ we can substitute in equation (12) directly from equations (19) and (20), to get

(21)
$$E = E_o - \frac{RT}{2F} \ln \frac{I}{I_d - I} - \frac{RT}{2F}$$

$$\ln \frac{\sqrt{(I_d - 2I)^2 + 16 I \kappa (I_d - I)} - (I_d - 2I)}{\sqrt{(I_d - 2I)^2 + 16 I \kappa (I_d - I)} + (I_d - 2I)}$$

This equation can be rewritten to give

(22)
$$E = E_o - \frac{RT}{2F} \ln \frac{I}{I_d - I} - \frac{RT}{2F}$$

$$\ln \frac{\sqrt{4\kappa I_d^2 - (4\kappa - I)(I_d - 2I)^2} - (I_d - 2I)}{\sqrt{4\kappa I_d^2 - (4\kappa - I)(I_d - 2I)^2} + (I_d - 2I)}$$

This equation for polarographic curves is the exact parallel of Michaelis and Schubert's equation

$$(23) \quad E = E_m - \frac{RT}{2F} \ln \frac{1 + \mu}{1 - \mu} - \frac{RT}{2F} \ln \frac{\sqrt{1 + (4\kappa - 1)(1 - \mu^2)} + \mu}{\sqrt{1 + (4\kappa - 1)(1 - \mu^2)} - \mu}$$

Here E is the observed potential; E_m is the potential which is obtained when $[Ox] = [Red]$. μ stands for $\frac{x}{a} - 1$, where a is the total molar amount of substance originally present in the oxidized form, and x is the amount of reducing agent added at a given stage of the titration, expressed in equivalents so that at the endpoint of the titration $x = 2a$. κ is the dismutation constant.

These equations demonstrate the symmetry of the potential around the midpoint of the curve. They are conveniently separated into two logarithmic components, the first of which holds if no intermediate form arises. In this case k is very small (κ very large) and the second logarithmic term vanishes. This second logarithmic term is the correction due to step formation, and the dismutation constant appears only in this term.

The shape of this function varies with the value of κ or k . This is demonstrated in FIGURES 1 and 2. If k is very small then the potential approaches the value

$$(24) \quad E = E_m - \frac{RT}{2F} \ln \frac{I}{I_a - I}$$

as it should be for a system with no intermediate step of reduction.

When $\kappa = 1/4$ or $k = 4$, then

$$(25) \quad E = E_m - \frac{RT}{F} \ln \frac{I}{I_a - I}$$

This formula represents a curve in which all ordinates are double the size of those in equation (24). This is the same curve as for a univalent reduction system. In the diagrams of FIGURES 1 to 6, the curves corresponding to these two conditions have been drawn with a heavy line to distinguish them from the remaining ones which represent k values of 1, 16, 100, and 1000.

FIGURES 1 to 6 have been prepared to give a clear demonstration of the type of curves which may be constructed on the basis of equation (22). FIGURES 1, 3 and 5 represent results obtained under ideal conditions in which applied potential values are plotted against three other functions. As has been stated before, these values are obtained directly from the polarogram *only* when the factor IR is negligible. All calculations are made for a temperature of 30° C. Since corrections for IR are tedious and it is often desirable to make an approxima-

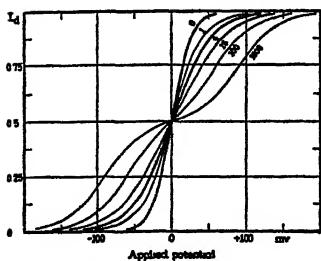


FIGURE 1.

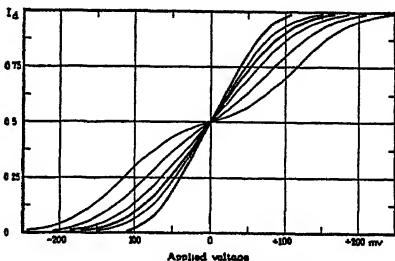


FIGURE 2.

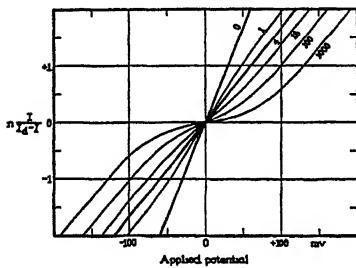


FIGURE 3.

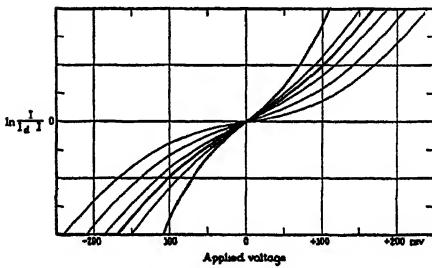


FIGURE 4.

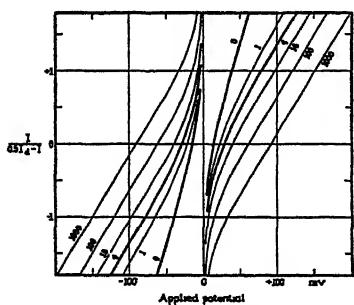


FIGURE 5.

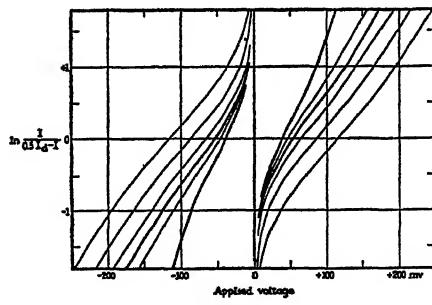


FIGURE 6.

tion by inspection of the polarographically obtained current-voltage curve, it seemed worthwhile to construct curves where the applied voltage is plotted as abscissa, neglecting the correction for IR . Such curves are presented in FIGURES 2, 4 and 6. They have been calculated on the basis that IR was equal to 100 mv. when the maximal diffusion current was reached. This corresponds to a current of 2×10^{-5} Amp. at a resistance of 5000 ohms, or to a current of 1×10^{-4} Amp. at a resistance of 1000 ohms. In polarographic practice, a

galvanometer with a maximum sensitivity of 2×10^{-9} Amp./mm./m. is often employed. Polarograms identical with the graphs of FIGURE 2 would then be obtained if 1/100 sensitivity were used and a wave 10 cm. high were plotted at a circuit-resistance of 5000 ohms. The same wave could, of course, also be obtained at a sensitivity of 1/500 with a resistance of 1000 ohms. A resistance of about 1000 ohms is found in simple polarographic circuits where a mercury layer at the bottom of the vessel is used as anode.⁶ In studies where the reference electrode is separated from the electrolysis vessel by agar bridges,³ the circuit resistances vary between 3000 and 6000 ohms.

Inspection of the graphs of FIGURES 1 and 2 shows that as k becomes greater than 16, the curves begin to separate into two distinct halves, each of which has a point of inflection. This seems to be just as marked in FIGURE 2 as in FIGURE 1. The difference between the half-wave potential, $E_{1/2}$ ($I = 0.5I_a$) and $E_{1/4}$ ($I = 0.25I_a$), or $E_{3/4}$ ($I = 0.75I_a$) has been called the index potential. If $k = 0$, the index potential is 0.0143 v. at 30° C. It is twice as large when $k = 4$, and it becomes equal to 0.03005 $\log k$, whenever $\log k > 2$.¹⁷

If these curves are analyzed by plotting $\ln(I/I_a - I)$ instead of the current I , the set of curves shown in FIGURES 3 and 4 is obtained. This type of analysis is common in polarographic work but as may be seen, the uncorrected curve for $k = 4$ (FIGURE 4) is almost identical with the ideal curve (FIGURE 3) where $k = 100$. Obviously, such analyses in order not to be misleading should be carried out only on current-voltage curves which had been corrected for IR . Furthermore, we find that when potentials are plotted, straight lines are obtained for k values of 0 and 4 (equations 24 and 25) with slopes of 30 and 60 mv. On the basis of previous experience in polarography, we would attribute these to a divalent and a univalent reduction, respectively. The latter conclusion is obviously wrong in this case; hence it is necessary to keep in mind the possibility of a semiquinone formation with a k value of 4, when a 60 mv. slope of $\ln(I/I_a - I)$ against potential is used as evidence for a univalent process.

Finally a set of curves has been prepared in FIGURES 5 and 6 in which only that fraction of the polarographic curves, which represents the addition of one equivalent in the reduction, has been considered in the analysis for $\ln(I/I_a - I)$. By plotting $\ln(I/0.5I_a - I)$ against applied potential or voltage, the two symmetrical sets of curves were obtained which approach the same potential asymptotically. The curves of FIGURE 5, furthermore, cross the zero line at the correspond-

¹⁷ Michaelis, L. Jour. Biol. Chem. 96: 703. 1932.

ing index potentials. It may be noted that all the curves of FIGURES 5 and 6 with k values of 1 or higher tend towards a 60 mv. slope, while the curve for $k = 0$ has a final slope of 30 mv. Unfortunately, the use of this difference for the determination of semiquinones will be limited because in practice the values at the beginning and at the end of the polarographic curves are much less reliable than those in the middle of the curves. However, similar analyses may prove of value in other polarographic work where different substances of unequal concentrations are reduced at almost the same potentials.

With these theoretical curves as ideal models, we can compare some experimental polarographic curves of semiquinones, to see how closely thermodynamic conditions are approached at the solution/electrode interface.

EXPERIMENTAL RESULTS

A number of compounds, for which Michaelis had established the existence of semiquinones, were studied polarographically with satisfactory results as long as the potentials fell within the polarographic range of $E_h + 0.6$ to -1.6 v.¹² For a demonstration of semiquinone formation α -oxyphenazine was found most suitable. Its semiquinone formation constant varies markedly within a convenient range of pH and potential.¹³ Two polarograms are reproduced in FIGURE 7 which were obtained by reducing α -oxyphenazine from air-free solutions* at different pH. It may be seen that the single polarographic wave at pH 8 and pH 6 breaks up into two distinct steps in 0.01 N and 0.1 N nitric acid** which are the farther apart the more acid the solution. To find the index potential in these cases, we have to measure the potential difference between $E_{1/2}$ and $E_{3/4}$ or between $E_{1/2}$ and $E_{1/4}$. However, a more easily determined value is the double index potential which is the difference between $E_{1/4}$ and $E_{3/4}$. Here we obtain directly from the polarogram the uncorrected values of 252 mv. for the 0.1 N HNO₃ and 158 mv. for the 0.01 N HNO₃ solution. The correction for IR is about 10 mv. so that we find 242 mv. and 148 mv. respectively for the double index potentials. This is in good agreement with the values of Michaelis¹³ who gives 244 mv. for pH 1 and 152 mv. for pH 2.

For pH 6 and 8 we find a corrected double index potential of 35 mv. instead of the theoretical 29 mv. This difference is still within the

¹² Michaelis, L. Jour. Biol. Chem. 92: 211. 1931.

* Since dissolved oxygen is reduced directly at the dropping mercury electrode, it is usually removed before the analysis by a stream of hydrogen or nitrogen gas.

** Hydrochloric acid cannot be used because it would limit the potential range.¹⁴

¹³ Michaelis, L. Chem. Rev. 16: 243. 1935.

experimental error but other deviations from the theoretical curves which will be discussed presently suggest that it may have a real significance.

If we analyze these experimental current-voltage curves as we did the theoretical curves, we obtain FIGURES 8 and 9. Here are plotted the results obtained with α -oxyphenazine in 0.1 N HNO₃ and 0.01 N HNO₃ (from FIGURE 7), in pH 4.3 buffer (from FIGURE 10), and in pH 6 buffer (from FIGURE 7), together with the theoretical potential

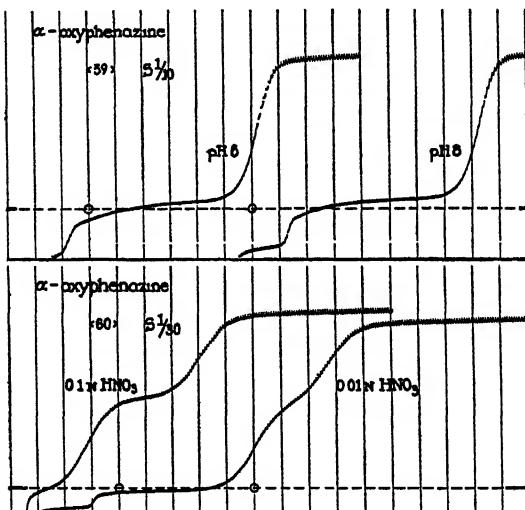


FIGURE 7. Polarograms of α -oxyphenazine in McIlvaine buffers of pH 6 and 8, and in 0.1 and 0.01 N HNO₃ solutions. For each curve, the potential corresponding to that of the calomel half-cell ($E_A = -0.244$ v.) is indicated by a circle.

curve for $k = 0$ (heavy line). Even if we apply a correction for IR (at most 20 mv.) we would still have a definite discrepancy between these experimental and the theoretical values. Especially the asymmetry of these curves with respect to the midpoint would still persist, since they are plotted with the midpoint as reference. Other cases of asymmetric polarographic curves have been explained by insufficient buffering²⁰ or by the irreversibility of the system,¹² but neither explanation is applicable here. However, some observations on very dilute solutions of α -oxyphenazine point to a possible explanation.

When the reducible substance was diluted in order to make the diffusion current as small as possible so that the IR correction would

* Müller, O. H. Paper presented at the Fall-meeting of the Am. Chem. Soc. Milwaukee, 1938.

be eliminated, it was found that the second step in the reduction was definitely smaller than the first. This is demonstrated in FIGURE 10 where α -oxyphenazine solutions in concentrations varying from 10^{-4} to 10^{-3} M were reduced at pH 1.3 and pH 4.3. The two steps observed at pH 1.3 are definitely unequal in the lower concentrations and approach each other in height only at the higher concentrations.

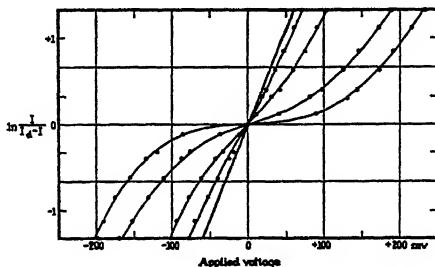


FIGURE 8. Experimental points from FIGURES 7 and 10 (0.1 and 0.01 N HNO₃; solutions and pH 4.3 and 6 McIlvaine buffers).

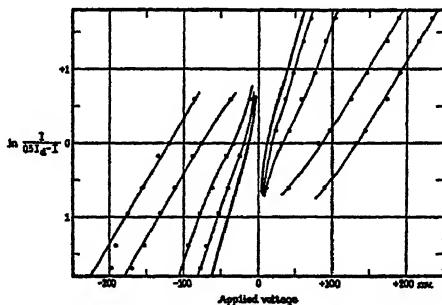


FIGURE 9. Experimental points from FIGURES 7 and 10 (0.1 and 0.01 N HNO₃; solutions and pH 4.3 and 6 McIlvaine buffers).

The explanation for this must be sought in the nature of the electrode reaction, and in the diffusion processes underlying this reaction. Experiments to clarify this point are still in progress. This discrepancy between theory and practice may have some advantages; for instance, in the curve due to the lowest concentration of α -oxyphenazine at pH 4.3 (FIGURE 10) a two-step reduction is indicated, while in the potentiometric determinations no separate steps can be observed at any pH > 3.3.

While we may conclude from this that the polarographic method at present is not refined enough for an accurate determination of the dismutation constants, we should not overlook some definite advan-

tages of this method over the potentiometric method. First of all we must point out the rapidity with which determinations can be made. Even if the results are not very accurate, it will be of value to get quick information about the existence of semiquinones and the range of pH in which they exist by inspection of the polarographic current-voltage curves or by measuring the double index potentials. To explore the possibilities for semiquinones over the range of pH 1 to pH 14 in steps of 1 pH unit, fourteen curves can be prepared on two or three polarograms in about two hours.

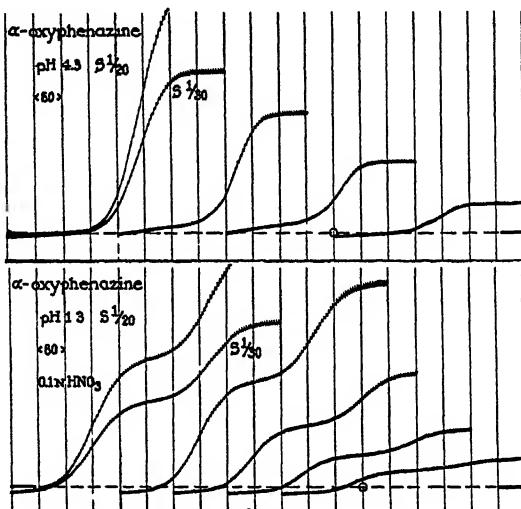


FIGURE 10. Polarograms showing the change in the waves as the concentration of α -oxyphenazine is increased from 0.0001 to 0.001 M.

The fact that the polarographic range extends to very negative potentials suggested another definite advantage of the polarographic over the potentiometric method for the study of semiquinones in the overvoltage range. Therefore, a solution of methyl viologen was polarographed at different pH values. Michaelis and Hill²¹ could study this compound only in the range of pH 9-13, because at lower pH the titration curves were incomplete due to the overlapping with the hydrogen potential. The polarographic method, however, gave satisfactory curves for this compound as low as pH 2.2. It may be seen from FIGURE 11 that the first wave has a half-wave potential (first arrow) which is constant over the whole range of pH and equal to $E_h = 0.441$ v. This is in good agreement with the results of

Michaelis and Hill who found a potential of -0.446 v. The conclusion of these authors that the potential of the first step in this reduction is independent of pH has thus been verified polarographically at very low pH values.

The second step in the reduction of methyl viologen could not be investigated by the potentiometric method because of drifts in potential due to secondary irreversible reactions.²¹ Since the dropping mercury electrode has proven to be especially suited for the study of such partially reversible reactions,^{11, 12} it seems likely that the second

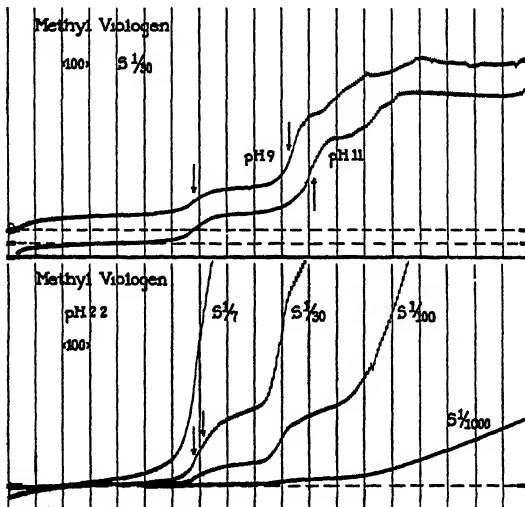


FIGURE 11. Polarograms of methyl viologen reduced at different pH.

step in the reduction of methyl viologen is represented by one of the subsequent curves found on the polarograms. The second set of waves in FIGURE 11 which has also been marked by arrows, is considered to be due to this reaction; however, since there is no simple relation between the height of the first and second set of waves, it is obvious that we have to deal here with a more complex mechanism than simple semiquinone formation. This second step in the reduction is not independent of pH and a 60 mv. shift in potential per pH unit may be observed as this wave gradually approaches the first one with decreasing pH.

This leads to a consideration of the reduction of organic compounds which are only partially reversible. Here the final reduction product

²¹ Michaelis, L., & Hill, E. S. Jour. Gen. Physiol. 16: 859. 1933.

cannot be oxidized at the dropping mercury electrode at the same potential as that at which the oxidized compound is reduced. Most polarographic studies of organic substances have been of this type. As has been shown by Müller²¹ and by Müller and Baumberger,²² the potentials obtained in these instances may be considered as due to a reversible step in a reaction which is, on the whole, irreversible. Therefore, the name "polarographic apparent reduction potential" (P. A. R. P.) has been suggested²¹ for those half-wave potentials which are not truly reversible, and which show dependence on pH, and demonstrate partial reversibility. The first products of the reaction must be in an equilibrium with the oxidant from which they are formed, but they are quickly removed from the solution by some secondary process. In this first reversible reaction, which may require

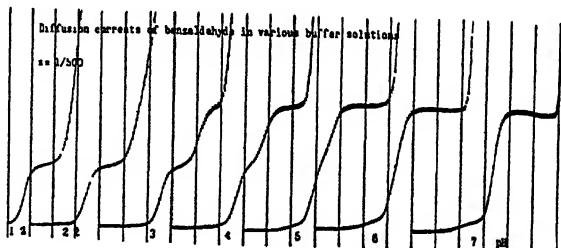


FIGURE 12. Polarogram of benzaldehyde reduced at different pH (from Tokuoka (23)).

two electrons, the formation of intermediate radicals is not impossible, and two examples may be cited from the literature which strongly suggest such a phenomenon in irreversible reductions. The first one is the reduction of benzaldehyde which was studied by Tokuoka.²³ A reproduction of his curves is shown in FIGURE 12. Another example is the reduction of benzophenone, studied by Schweitzer and Laqueur.²⁴ In both instances the single wave obtained in the more alkaline solutions separates into two equal halves upon acidification of the solution.

The simple ratio in these stepwise reductions need not necessarily be unity in all cases. For instance, when picric acid is reduced in solutions which are more acid than pH 4, two waves are obtained with a relative height of 3 : 2. As the solution becomes more alkaline, the first large wave breaks up into two components of a fixed ratio of 1 : 2,

²¹ Müller, O. H., & Baumberger, J. P. *Jour. Am. Chem. Soc.* 61: 590. 1939.

²² Tokuoka, M. *Collection Czechoslov. Chem. Commun.* 7: 392. 1935.

²³ Schweitzer, H., & Laqueur, E. *Rec. trav. chim.* 55: 959. 1936.

so that at pH 8 the reduction goes on in three steps with a ratio of 1 : 2 : 2.¹² Similar results have been obtained with other nitrated phenols.

These reductions in steps of fixed ratio suggest that the reactions are determined by the addition of a fixed number of electrons to each reducible molecule. This differs from another type of stepwise reduction in which the steps are not in a fixed ratio. The two types of reactions should not be confused because in the latter case a fraction of the reacting substances undergoes a change in structure before it is reduced, due to pH changes in the solution. For instance, Müller and Baumberger have shown²² that in the keto-enol tautomerism of pyruvic acid and in the polymerization of this compound, which are governed by the pH of the solution, different waves are obtained for each component of the solution. The relative heights of these waves are not in a fixed ratio, but one wave gradually increases as the other diminishes with changes in pH.

A word needs to be said about another type of two-step reduction which has been observed in potentiometric work, if the semiquinone radicals combine to form a quinhydrone or meriquinone.⁵ This may be distinguished from that due to a simple semiquinone by changing the concentrations of the reactants. In reactions at the dropping mercury electrode, in which the semiquinone radical is produced at the electrode interface only during the life-time of a drop, one should expect no difference unless the process of dimerization is as fast as the electrode reaction. Also the solution would have to be more concentrated than in the usual analyses. Nevertheless, there exists the possibility of a polarographic study of meriquinones when they have been prepared by the addition of a reducing agent to a fairly concentrated solution.

In conclusion it may be stated that the polarographic method may be used satisfactorily for orientative experiments to determine the possibility of semiquinone formation in the reduction or oxidation of reversible organic oxidation-reduction systems. With certain reservations, an extension of the formulas and equations developed for potentiometric studies of such systems is possible for polarographic work. The potential range which can be studied under proper conditions is from $E_h + 0.6$ to -1.6 v. The most outstanding advantage of the polarographic method is its applicability to potential measurements in the overvoltage range and its suitability for studies of systems, which are only partially reversible. Even in these cases the formation of intermediate radicals can be demonstrated polarographically.

THE ANALOGY BETWEEN TWO-STEP OXIDATION AND TWO-STEP IONIZATION

By MAXWELL SCHUBERT

From the Laboratories of The Rockefeller Institute for Medical Research, New York

Protons and electrons are the most mobile and active participants of chemical changes. The proton takes part in acid-base equilibria. It is the product of the dissociation of an acid:



The electron takes part in oxidation-reduction equilibria. It is the product of the dissociation of a reductant:



Acid-base equilibria are usually treated by the mass law equation:

$$K_{Ac} = \frac{[H^+] [A^-]}{[HA]} \quad (3)$$

Then for oxidation-reduction equilibria we might have expected a parallel equation:

$$K_{Ox} = \frac{[E] [A]}{[A^-]} \quad (4)$$

But while the fiction of free protons in solution implied in equation (3) was readily accepted, the fiction of free electrons in solution has never been much used. That is to say, acetic acid was believed to dissociate in solution into protons and acetate ions and "hydrogen ion concentration" is an accepted phrase. But ferrous ion has never seriously been considered to dissociate into ferric ions and free electrons to an extent to render "electron concentration" an acceptable phrase. Actually, of course, neither electrons nor protons occur in solution except in the most minute amounts. The reason acids appear to dissociate according to (1) depends on the accidental fact that our common solvents as water and alcohols are proton acceptors. These same solvents do not at all readily accept electrons so the dissociation (2) does not occur in water and has never been considered as a real equilibrium. If it happened for example that we used stannic chloride as a solvent we might speak of a progressive dissociation of ferrous ion into ferric ion and an electron as we diluted with the electron accepting solvent. And if we measured the separate ionic species we might find

an equation as (4) to be realized. In this case "electron concentration" as meaning free electrons would be just as much of a fiction as "proton concentration" is in aqueous solutions.

In studies of oxidation-reduction equilibria this difficulty has been avoided by describing the process with the equation:

$$E = E_{Ox} + \frac{RT}{F} \ln \frac{[A]}{[A^-]} \quad (5)$$

where E_{Ox} is a constant for a given system, the normal oxidation potential. A similar treatment of acid-base equilibria has been advocated by Schwarzenbach¹ who has recommended the use of a normal acidity potential, E_{Ac} , defined by the equation:

$$E = E_{Ac} + \frac{RT}{F} \ln \frac{[HA]}{[A^-]} \quad (6)$$

in place of the acid dissociation constant. Then equations (5) and (6) in which electron and proton concentrations are avoided can be used in place of (3) and (4). In this way the fundamental similarity of one-step acid ionization and one-step oxidation is made apparent.

If we make a formal comparison between reversible two-step oxidation reactions and reversible two-step proton ionizations we have simply the pairs of equations:



and:



Each of these sets of reactions has been extensively studied but again, as for the one-step cases, the underlying theory has been developed quite differently. Yet if we look at these changes and realize that in each case simple mass law relations are assumed to hold, there is no reason apparent why there should not be complete identity in the equations developed to describe the two sets of changes. Actually this identity in the equations can easily be shown.

The theory of acid titrations has been developed in terms of the two ionization constants characterizing the reactions (7) and (8).

$$K_1 = \frac{[H^+] [HA^-]}{[H_2A]} \quad (11)$$

¹ Schwarzenbach, G. *Helv. chim. Act.* 13: 874. 1930.

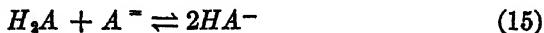
$$K_2 = \frac{[H^+] [A^-]}{[HA^-]} \quad (12)$$

On the other hand the theory of reversible oxidations has been developed by Michaelis in terms of the semiquinone formation constant of the reaction (13) which is implied in the reactions (9) and (10).



$$k = \frac{[A^-]^2}{[A] [A^-]} \quad (14)$$

We can set up equations analogous to (13) and (14) for the case of acid ionization as follows:



$$k' = \frac{[HA^-]^2}{[H_2A] [A^-]} \quad (16)$$

comparing (11) and (12) with (16) we have finally:

$$k' = \frac{K_1}{K_2} \quad (17)$$

Equation (14) describing oxidation reactions and equation (16) describing acid ionization are equivalent and equation (17) shows that the analogue of the semiquinone formation constant is the ratio of the two ionization constants of an acid.

This formal comparison would make it appear that the ionization of dibasic acids which can lose two protons and the oxidation of reductants which can lose two electrons should be very similar processes. Actually experience has shown the two phenomena to be so different that, as pointed out above, their theoretical developments have proceeded along different lines. The study of dibasic acids showed that in general K_1 is much greater than K_2 , so that k' of equation (17) is generally between 10 and 100,000. In fact for a symmetrical dibasic acid the lower limit of k' has been shown on theoretical grounds to be 4, but in practice very few substances have been found which approach this limit. On the other hand studies on reversible oxidation-reduction systems until recently showed that for all organic systems k was zero. In fact so certainly and tacitly was k regarded as being a non-existing quantity that no one appears even to have mentioned its non-existence. The physical interpretation of this is that after one electron has been removed from the substance A^- , as in equation (9), then the resulting molecule A^- ejects a second electron

with much greater ease as in equation (10). So the molecule A^- was allowed no particular existence. The extensive work of Michaelis, however, has raised k from nothing to something and has given to A^- a recognition of its existence.

This extension of the range of the recognized values of k in oxidation-reduction systems makes one wonder whether an extension of the values of k' for acid systems is not also possible. Such an extension would be in the opposite direction however, and the question posed is this: Can k' ever become less than the theoretically derived quantity 4 and can it ever become less than unity, possibly even approaching zero as k used to be supposed to do?

It is the purpose of the present discussion to examine the possibilities for such a phenomenon and the conditions which might make it possible. There will also be pointed out some fundamental differences between proton addition and electron addition. In the discussion, the term acid will be used in the more general sense including not only substances which dissociate protons but substances which associate hydroxyl ions. For example, potassium phenanthrene quinone sulfonate is an acid whose pK is about 11.

First, let us consider some of the more foolish consequences of the supposition that k' for a simple dibasic acid could be less than one. Then in equation (17) we would have K_2 greater than K_1 . But, if K_2 is greater than K_1 then the second proton would ionize at a pH more acid than that at which the first would ionize. Then what was called the second ionization would really be the first. This can not be what we are looking for.

In order to find out if there is any sense to our problem, let us examine the case of such reversible oxidation reduction systems as have values of k less than one. In this case, the second electron is ejected more readily than the first. Then why does it not come off first? The answer must be found in a more detailed search into the electronic structure of the molecules involved. Suppose A^- of equation (9) to be a simple organic reductant with an even number of electrons. Then A^- must have an odd number of electrons. In the case of a reductant such as ethyl alcohol for example, loss of a single electron would give a radical at an oxidation level between ethyl alcohol and acetaldehyde. In such an odd electron molecule there are not many positions which the odd electron can assume, so stabilization of the molecule by formation of a resonance system is not possible. The radical is unstable and has a great tendency to lose a second electron to give the more stable electronic configuration of acetalde-

hyde. In such a case k might easily be expected to be less than one, though in the case of ethyl alcohol this constant cannot be measured. On the other hand, if by removal of a single electron from a reductant, there is produced an odd electron molecule in which there is great stabilization due to resonance, then, this molecule may have a much smaller tendency to lose a second electron than the original molecule. Then k will be greater than one and we have the case where a stable semiquinone is formed. In any case, the molecules A^- and $A^{\cdot -}$ have a similar distribution of atomic nuclei but an altogether different electronic structure. They are really quite different molecules and their reduction potentials may be either very different or very similar and k may vary from values close to zero to very large values.

Now let us examine the types of dibasic organic acids whose titration curves have been studied. These are largely carboxylic, phenolic or ammonium types. In such cases removal of a single proton leaves both the nuclear and the electronic arrangement of the molecule essentially intact. Then the removal of a second proton is from essentially the same molecule and is affected only insofar as the statistical and electrostatic conditions in the molecule are different. So, while the molecules A^- and $A^{\cdot -}$ of oxidation titrations have generally very different electronic structures the molecules H_2A and HA^- of acid titrations have generally quite similar structures.

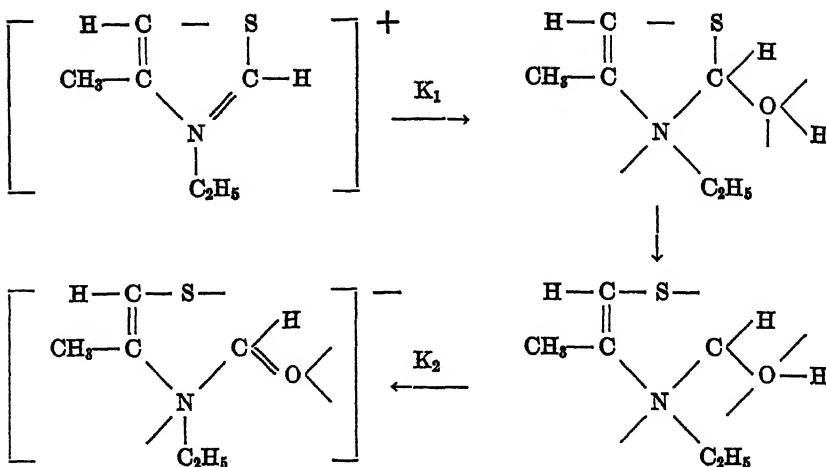
By analogy then, if we are to find a substance, for which even a possibility exists that k' might be less than one, we should look for a substance such that the first ionization step whether addition or loss of a proton or a hydroxyl ion brings about a fundamental rearrangement of the electronic structure of the molecule of such a nature that the new molecule formed undergoes a second such change more readily.

To search for such a substance we have no guiding principle but fortunately a case has already been described which furnishes an example. This occurs among certain thiazolium and benzothiazolium salts of which thiamine or vitamin B_1 is one example. The nature of the changes which occur on treating aqueous solutions of these substances with two equivalents of alkali was worked out by Mills, Clark and Aeschlimann.² A few years ago Williams and Ruehle³ published some acid base titration curves of a few of these substances. Clarke and Gurin⁴ studied some further examples. As an example let us take 3 ethyl 4 methyl thiazolium iodide. On adding alkali to an aqueous solution of this substance the following changes occur:

² Mills, W. H., Clark, L. M. & Aeschlimann, J. A. *Jour. Chem. Soc.* 123: 2353. 1923.

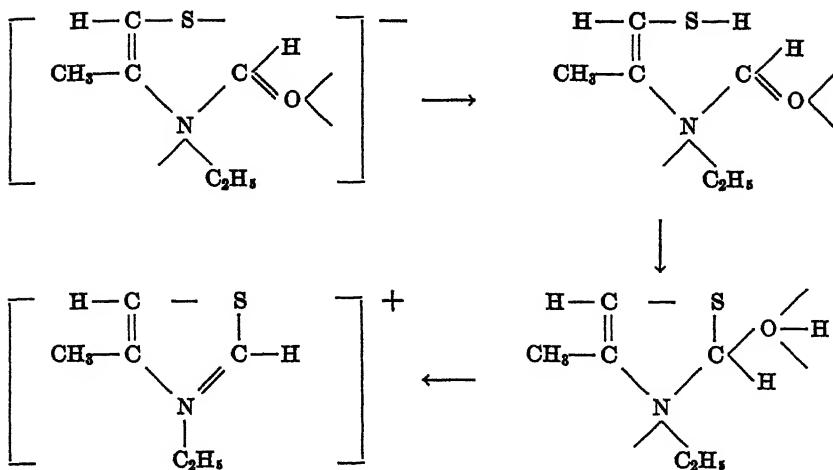
³ Williams, R. E., & Ruehle, A. E. *Jour. Amer. Chem. Soc.* 57: 1876. 1935.

⁴ Clarke, H. T., & Gurin, S. *Jour. Amer. Chem. Soc.* 57: 1876. 1935.



The first step is the association of a hydroxyl ion and the formation of a molecule which would be called a pseudo base. This has a resemblance to the aldehyde thiol addition compounds which at the pH now existing are known to be dissociated, so in the next step the dissociation occurs, resulting in a compound which is seen to be the acid form of a formylated amine and at the pH required to bring about the first step of the reaction chain it must lose a proton instantly. That is K_2 is greater than K_1 but the second step of ionization could not occur before the first had taken place and been followed by the rearrangement. The titration curve bears out this interpretation fully. If we apply to this curve the type of analysis used in oxidative titrations we find an index potential of about 14 millivolts indicating the simultaneous absorption of two hydroxyl ions the first of which attaches to the thiazolium ion and the second is absorbed by the ejected proton.

Although this reaction is reversible if we consider the change from one end of the reaction chain to the other, it is impossible that in detail this should be so. Consider the negative ion finally produced by titration with alkali. On making its solution acid a proton would attach first, not to the oxygen of the formyl group, but to the mercaptide ion which has a pK of the order of 10. The reverse sequence of reactions would be:



This brings out an interesting point in the fact that though the overall reaction is reversed the intermediate steps are not the same in titrating with acid or with alkali. Oxidation reduction equilibria so far studied by potentiometric methods on the other hand show complete identity of the oxidative and reductive curves. Intimately associated with this contrast is another. In the acid or base titrations of the thiazolium systems discussed above equilibria are established slowly. This is made apparent in the titration by the rapidly drifting potentials observed after each addition of alkali or acid. A wait of at least fifteen minutes is required before this potential drift approaches a value which may roughly be considered near the equilibrium value. The reason for this lies in the slowness with which the rearrangement reaction takes place. In potentiometrically studied oxidation reduction titrations equilibria are generally established quickly, almost instantaneously. This shows that the fundamental rearrangement of electron distribution which must take place and which is generally accompanied by a color change occurs very rapidly.

This contrasting behavior of acid base and oxidation reduction systems has at its base the very different mobilities of electrons as against atomic nuclei. When a proton or a hydroxyl ion is attached to or detached from a molecule, this occurs at a definitely definable position. The rearrangement which follows this is a rearrangement of atomic nuclei within the molecule and this may have the slowness of organic reactions. When an electron is attached to or detached from a molecule we generally do not know in detail at what point in the molecule this occurs or even whether such knowledge is possible. All we

do know is that within the framework of the atomic nuclei the electrons are mobile enough so that their redistribution occurs with such rapidity so that as far as our measuring instruments are concerned it is instantaneous.

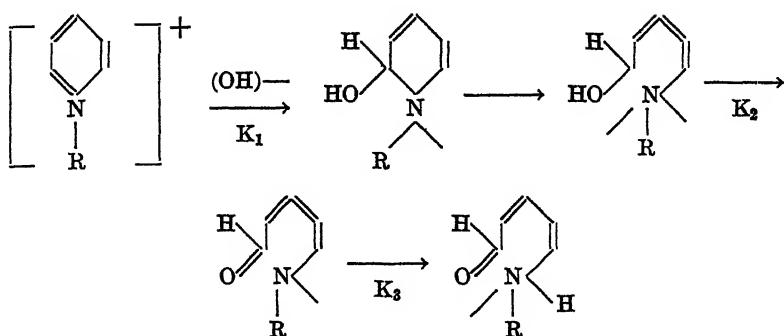
While we are discussing these contrasts between oxidation-reduction reactions as opposed to acid base reactions it will be well to bring in at this point a practical distinction which must be made in the constants. Potentiometric oxidation-reduction titrations are always run in solvents such as water, alcohol, acetic acid, pyridine or mixtures of such solvents, all of which are proton donors or acceptors. The effect of this is, that together with the equilibria (9) and (10) which represent the elementary oxidation reaction there are always acid-base equilibria in play simultaneously. The molecules A^- , A^- and A will in general exist at different ionization levels. Then the values of k actually measured will be dependent on the acidity of the medium in which the measurement is made. It is therefore called a composite constant to distinguish it from the simple or elementary constant defined in (14). In contrast to this, acid base reactions are always studied in solvents which are not electron donors or acceptors under the conditions used. That is, the molecules H_2A , HA^- and A^- are all at the same oxidation level. The values of k' are thus really characteristic of the elementary reactions (7) and (8). Values of k and k' cannot be directly compared but k must first be converted to some elementary value, that is a value such as in (14) where all the oxidation levels are at the same ionization level. In this there will, of course, be some element of arbitrariness depending on what ionization level is taken as fundamental.

As an example of the type sought, the thiazolium salts are excellent. They form a sort of reversible system in which two protons or two hydroxyl groups are apparently simultaneously absorbed. It is a double step acid base titration with no separation of the steps and with a value of k' of zero. It is of interest that the related compounds in which the 2 position is substituted or in which the 3 position is not substituted do not show this behavior.

But the search for other examples of this sort has not yielded such good ones. A case which we are now investigating is among pyridinium salts and another may be among benziminazolium salts. For example, among pyridinium salts the 2, 4 dinitrophenyl pyridinium chloride has been extensively studied.⁵ On making an aqueous solution of this substance alkaline a series of changes occurs which can be

⁵ Zincke, T. Ann. 330: 361. 1904.

represented by the following scheme, where R is the dinitrophenyl group:



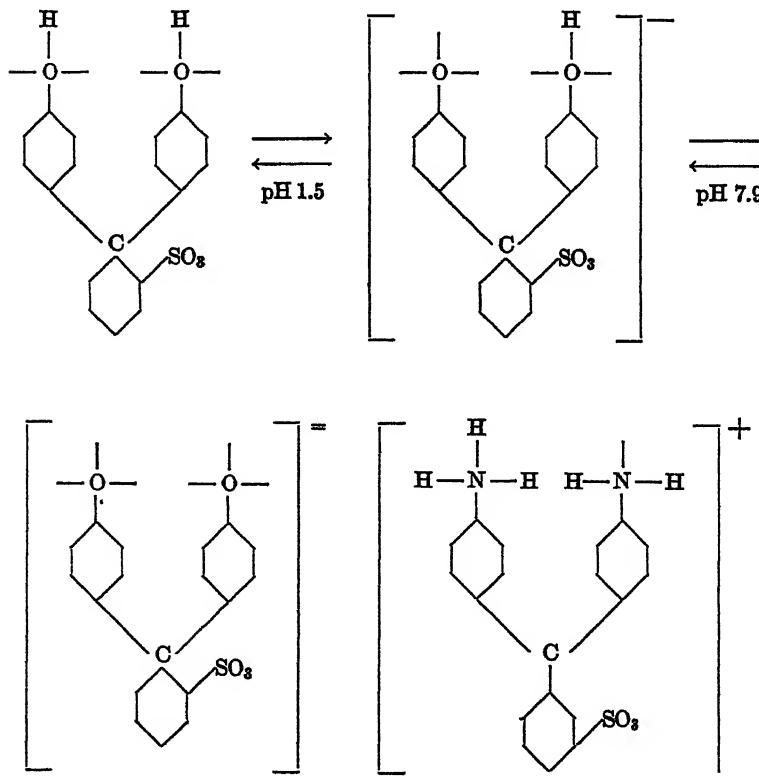
The principles involved are the same as have been described for the thiazolium salts but in this case the reaction in aqueous solution is reversible only in small part. The final product of the alkaline titration, if isolated and titrated back in acetic acid with hydrochloric acid gives the original pyridinium salt almost quantitatively. Another difference in this case is that the second step at K^2 where a second hydroxyl ion would be absorbed by the ejected proton is masked by a further overlapping step at K^3 where a proton is absorbed.

The principles outlined above show that double step acid base titrations with values of k' approaching zero do exist. Such systems are too little known at present and it is hoped that more examples will soon be brought to notice. There are some essential differences between such double step acid-base systems and redox systems. These differences seem to depend on the slow rearrangement of atomic nuclei to form the intermediate. But there seems to be no reason why k' may not assume all possible positive values.

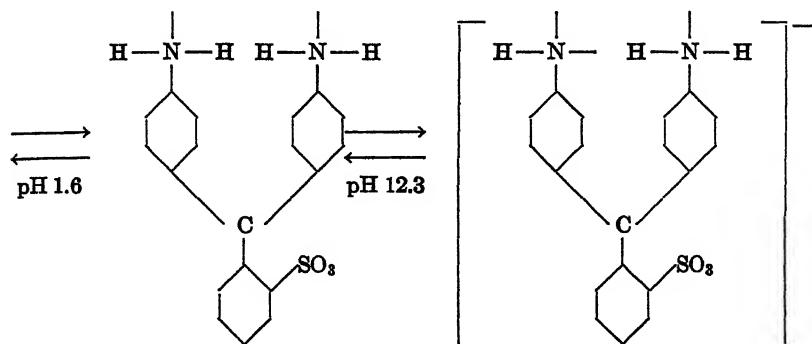
There is another principle which it may be worth discussing here and that may be an important factor in modifying the ratio of two successive acidic constants though sometimes it can operate to increase this ratio and to decrease it. In the analysis above of the conditions likely to lead to closer approximation between successive ionization constants a rearrangement of the molecule to form an intermediate of different electronic configuration was pointed out as necessary. In the examples cited a change in nuclear configuration occurred as well, carrying in its train of course a change in electronic distribution. But it is possible to have a rather deep-seated change in electronic structure as a result of an ionization step without a change in nuclear distribu-

tion. This occurs with many indicators at their turning points where the change in electron pattern causing the color change is not necessarily accompanied by a change in the nuclear pattern. A particular case of this sort is that in which successive ionizations progress through symmetrical and unsymmetrical structures. Such systems have been studied by Schwarzenbach who has clearly stated the underlying principles for these cases.⁶

Symmetrical resonance systems have a greater stability than unsymmetrical systems and so individual constants will tend to be shifted in such ways that symmetrical systems have a greater range of existence over the pH scale. The systems studied were largely the phthaleins and as examples may be taken phenol red and aniline sulfonaphthalein.



⁶ Schwarzenbach, G., Brandenberger, M., Ott, G. H., & Hagger, O. *Helv. chim. Act.* 20: 490. 1937. Schwarzenbach, G., & Hagger, O. *Helv. chim. Act.* 20: 1591. 1937.



Here the symmetrical form of the aniline sulfone phthalein exists over a range of over ten pH units while the unsymmetrical ion of phenol red has an existence over only about six units. Phenol-phthalein has been frequently studied and the most reliable measurements indicate a ratio of the two constants to be just about four.⁷ This would mean that the tendency to form a symmetrical resonance system is so great that the statistical limit of the value of k' is reached although the distance between the ionizing groups would lead one to expect a value of over 7 for k' . However, the case of phenolphthalein is complicated by lactone formation and may be open to other interpretations.

One final case may be mentioned here for which no adequate interpretation can be given.⁸ For phloroglucinol there have recently been published data which appear reliable enough at least as to order of magnitude showing a value of k' of 2.7. Phloroglucinol is a symmetrical tribasic acid so the statistical limiting ratio of the first to the second ionization constants is three. But the acid groups are quite close in phloroglucinol so a much larger ratio would be expected, perhaps of the order of 150 found for resorcinol. Any suggestion as to how to account for the unexpectedly low value will be welcomed.

DISCUSSION BY DR. G. W. WHELAND

Ingold, *et al.* concluded from the abnormally high ratios of the first and second dissociation constants of the alkyl malonic, succinic, and glutaric acids that the substituents had decreased the $C-C-C$ valence angles in the chain and hence the distance between the carboxyl groups. Westheimer and Shookhoff, however, have given a different interpretation of the phenomenon. Following Kirkwood and

⁷ Thiel, A., & Diehl, E. Sitzungsber. Beförd. gesamt. Naturwiss. Marburg 62: 472. 1927.

⁸ Abichandani, C. T., & Jatkar, S. K. K. Jour. Indian Inst. Sci. 21A: 417. 1938.

Westheimer, they consider the organic molecule in aqueous solution as a cavity of low dielectric constant immersed in a medium of high dielectric constant. The electrostatic potential at one point in this cavity due to a charge e at another point in the cavity can be represented as $V = e/D_E r$, where D_E is an "effective dielectric constant" and r is the distance between the points. The value of D_E depends not only upon the dielectric constants of the medium and of the molecule but also upon the size and shape of the cavity. It is evident, for example, that in the once ionized alkyl malonic acids the potential at the position of the remaining ionizable hydrogen atom due to the negative charge on the carboxyl ion will be related to a comparatively small effective dielectric constant. This is because the lines of force must largely pass through the molecule itself, that is, through a region of low dielectric constant. In the unsubstituted malonic acid, on the other hand, D_E is larger, because more of the lines of force have to pass through the aqueous medium. Thus the difference in the ratios of the ionization constants of the substituted and unsubstituted acids is due to variations not in the distance r , as assumed by Ingold, but in D_E . Quantitative calculations based upon this interpretation have led to the result that the distance between the two carboxyl groups is very nearly independent of the presence or absence of substituents.

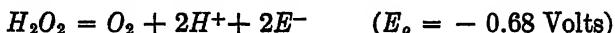
THE FREE ENERGY OF O₂⁻ IN RELATION TO THE SLOWNESS OF OXYGEN REACTIONS

BY MANUEL H. GORIN

From the Biological Laboratory, Cold Spring Harbor, New York

It is well known that molecular oxygen reacts extremely slowly with most reducing agents in aqueous solution in spite of favorable overall thermodynamic conditions. Furthermore, no rapidly reversible equilibrium in simple systems containing oxygen, hydrogen peroxide and water and another oxidation-reduction couple has ever been isolated in aqueous solution. Also, many oxygen reactions in aqueous solution appear to have chain characteristics. They are catalysed by minute amounts of active substances and the catalysis may be positive or negative (inhibitory). Induced¹ reactions involving oxygen have been isolated and carefully studied.

To explain the exceptional behavior of the oxygen molecule it has been postulated that the slowness with which it reacts is in some way connected with its high heat of dissociation into atoms. That this idea, alone, is insufficient to account for its behavior is evident. For it has been established that peroxides may be formed as intermediates in the reduction of molecular oxygen. Therefore, since in the formation of hydrogen peroxide the O-O bond is not broken and also since the potential² of the half reaction



is favorable, no direct connection appears to exist between the high heat of dissociation and the unreactivity of the oxygen molecule.

The recent work of Michaelis³ on the two step oxidation (one electron exchanges) of the quinones to hydroquinones and the existence of intermediate free radicals, the semi-quinones, has led him to postulate that in the main oxidation-reduction reactions take place by exchanging only one electron at a time. This theory is difficult to refute, for it is impossible to demonstrate by known methods that a given reaction proceeds by the transfer of two electrons simultaneously, even though no signs of step-wise reactions are detectable. On the

¹ Bray, W., & Ramsey, J. B. *Jour. Am. Chem. Soc.* **55**: 2279. 1933.

² Latimer, W. "The Oxidation States of the Elements," Prentice Hall, Inc., New York. 1938.

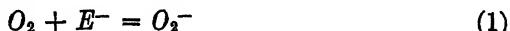
³ Michaelis, L. *Chem. Reviews* **16**: 244. 1935.

Michaelis, L., & Fletcher, E. S. *Jour. Am. Chem. Soc.* **59**: 2460. 1937.

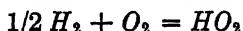
Michaelis, L., Boeker, G. F., & Reber, R. K. *Jour. Am. Chem. Soc.* **60**: 202, 214. 1938.

other hand, it is sufficiently established to be used as a working hypothesis and, as will be shown, suggests an explanation of the behavior of molecular oxygen.

Using the one-electron hypothesis of Michaelis³ the first step in the reduction of oxygen might be

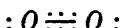


To estimate the potential of the above half-reaction it is necessary to obtain the change in free energy of the reaction

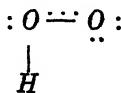


While several indirect estimates⁴ have been made no direct thermodynamic evaluation by spectroscopic or other means of the standard free energy of HO_2 are available. However, the ideas of Pauling⁵ on the approximate additivity and constancy of bond energies can be used to estimate ΔH for the formation of HO_2 . Since ΔH and ΔF will not be expected to be widely different and the free energy of hydration is small a fairly good approximation for ΔF might be obtained by applying these ideas.

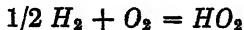
Pauling⁶ gives for the structure of oxygen the following:



Furthermore, added stability estimated by Wheland⁶ to amount to about 20.0 kcal. apparently results from the coupling of the two three-electron bonds. It might be fairly accurate, then, to divide the heat of dissociation of the oxygen molecule, 118 kcal. per mol., thus: single $O-O$ bond 35 kcal., coupling between two three-electron bonds 20 kcal. and, therefore, 63 kcal. remain for the two three-electron bonds. The structure of HO_2 is probably



with one single $O-O$ bond (35 kcal.) one three-electron bond ($63/2 = 31.5$ kcal.) and one $O-H$ bond (110 kcal.). It follows that for



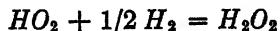
⁴ Bray, W. Jour. Am. Chem. Soc. 60: 82. 1938.

Weiss, J. Trans. Faraday Soc. 31: 668. 1935.

⁵ Pauling, L. "The Nature of the Chemical Bond." Cornell Univ. Press, Ithaca. 1939.

⁶ See Reference 5 page 253.

$\Delta H_o = - (110 + 35 + 31.5) + 51.5 + (35 + 63 + 20) = - 7 \text{ kcal.}$
 Similarly, for



$\Delta H_o = - (220 + 35) + 51.5 + (35 + 110 + 31.5) = - 27 \text{ kcal.}$

For



and

$$\Delta F_{(aq)} = - 36.5$$

or

$$\Delta F_{(aq)} = \Delta H_o + 2.5$$

Similarly, for,



and

$$\Delta F^o = - 56.5$$

or

$$\Delta F^o = \Delta H_o + 1.5$$

By analogy, assuming that the quantity $\Delta F_{(aq)} - \Delta H_o$ will be about the same amount greater for HO_2 compared with H_2O_2 as it is for H_2O_2 compared with H_2O , it comes out to be about 3.5 kcal. Therefore, ΔF^o for $HO_2 = - 7.0 + 3.5 = - 3.5 \text{ kcal. per mol.}$

For the reaction



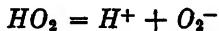
$$\Delta F^o = - 37 + 56 = 19$$

While for



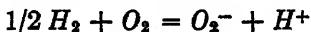
$$\Delta F^o = 31.5 - 15.5 = 16.0$$

If the trend were to continue ΔF^o for



would be about 3 kcal. less than for reaction 2, or 13 kcal. The value, 13.0 kcal. roughly agrees with that suggested by Bray⁴ from the similarity between HOO and $HOCl$, hypochlorous acid.

To summarize, for the reaction



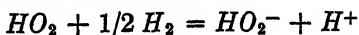
taking $F^o_{H^+} = 0$, $\Delta F^o = - (- 3.5 + 13) = - 9.5 \text{ kcal. per mol.}$
 Or, finally, for the half reaction



The next step in the reduction probably is (O_2^- would be expected to capture proton instantaneously)



E_o for reaction 3 may be obtained as follows; for



$$\Delta F^o = F^o_{HO_2^-} - F^o_{HO_2} = - 15.5 + 3.5 = - 12 \text{ kcal.}$$

Therefore

$$E_o = - 0.50 \text{ Volts}$$

Since oxygen is such a weak oxidizing agent for the first step in which, in addition, considerable activation energy might be required, the behavior of oxygen outlined in the first paragraph might be expected. Reducing agents powerful enough to initiate rapidly the reduction of oxygen would be so completely oxidized at equilibrium that no detectable amounts would remain. Furthermore, since the intermediate HO_2 is a much more powerful oxidizing agent than oxygen and, in addition, probably requires less activation energy for the capture of an electron it might be expected to have the property of an active intermediate, which under certain circumstances, could take part in the formation of reaction chains. Similarly, induced reactions might be expected, but the argument in this case is more involved.

Certain conditions might serve to decrease both the free energy and the energy of activation required to initiate the reduction of the oxygen molecule. For instance, if a proton is captured as well as an electron ΔF^o decreases to $- 3.5$ kcal. However, by analogy to the case of two electrons, the probability of electron and proton capture taking place simultaneously should be very small.

Another possibility that must be considered in certain cases is that the reduction of the oxygen molecule might take place adiabatically (simultaneous breaking of old and forming of new chemical bonds) rather than by electron capture. Here of course an entirely different approach must be made. Certainly, such processes are of importance in the gas phase. In the main, however, the reactions of gaseous oxygen are also slow and subject to formation of reaction chains.

There is one possibility which appears to be the most important in providing mechanisms for many oxygen reactions. If the O_2 molecule enters into complex formation with another substance prior to its reduction the free energy relationships in the various steps of the reduction as well as the energy of activation required may be quite

different. In this direction may lie the explanation of the importance of copper as a catalyst for a great many oxygen reactions. Since cuprous ion forms complexes with carbon monoxide, it is not altogether unlikely that it does so with oxygen.

For more specific oxidation-reductions, the importance of the porphyrines as enzymes might also be closely associated with the formation of complexes with molecular oxygen.

The semiquinone radicals of the different oxidation-reduction systems are in general dissolved in water and thus, in order to measure their concentration by the para-hydrogen conversion, it is essential to measure the conversion of the dissolved hydrogen in such a solution. Experimentally this is done by shaking para-hydrogen with the solution of the paramagnetic substance in a closed vessel and by taking into account that only a certain part of the hydrogen—namely, that which is dissolved—is undergoing conversion.

If we denote the observed rate of conversion with C_{obs} , then the true rate of the conversion (C_z), due to the paramagnetic radical, will be given by the equation

$$C_z = [X] C_o = C_{obs} \cdot \frac{v_g + v_e \alpha}{v_e} \quad (2)$$

where v_g and v_e are the volumes of the free gas space and liquid respectively, and α is the solubility coefficient of hydrogen in the solution in question.

The rate constant C_{obs} may be derived from the half life duration time of the conversion

$$t_{obs} = \frac{\ln 2}{C_{obs}}$$

and is experimentally determined by following the change of the p-H₂ concentration with time during the conversion. The p-H₂ concentration is measured by the heat conductivity at low temperatures, the heat conductivities of p-H₂ and o-H₂ differing from each other by several per cent at 160° abs. In order to estimate accurately the concentration of the paramagnetic radical from the rate of conversion some corrections have to be made in equation (2).

First of all, the conversion due to the pure solvent (in most cases of oxidation-reduction systems, water) must be deducted from the observed conversion. It was found that diamagnetic liquids also catalyze (very slowly) the para-ortho conversion, the rate constant being $3 \cdot 10^{-7}$ to $3 \cdot 10^{-6}$ mole · liter⁻¹ · sec.⁻¹. In these cases the conversion is due either to the nuclear paramagnetism of the molecule (e. g. in case of water) or to some weak magnetic moments generated by the rotation of the molecule (e. g., with CS₂).

In aqueous solution the contribution of the solvent to the observed conversion amounts to

$$C_{H_2O} = 1.13 \cdot 10^{-6} \cdot 55.4 = 6.26 \cdot 10^{-5} \text{ mole} \cdot \text{liter}^{-1} \cdot \text{sec.}^{-1}$$

and instead of equation (2) we obtain

$$C_{\Sigma} = C_o [X] + C_{H_2O} = C_{obs} \frac{v_e \alpha + v_g}{\alpha v_e} \quad (2a)$$

The conversion caused by the water becomes important, when the concentration of the paramagnetic radical is smaller than $6 \cdot 10^{-3}$ mole/liter (correction ~ 10 per cent).

The second correction to equation (2) relates to the change of the solubility of hydrogen in the presence of the oxidation-reduction system. In general, by increasing the amount of dissolved material the solubility will be diminished and the use of α , as valid for pure water, would lead to an apparently too low value for C_{Σ} and thus for $[X]$. Therefore, especially in concentrated solutions, this effect must be taken into account.

In order to estimate the concentration of a semiquinone radical in an oxidation-reduction system it is best to investigate the system at the midpoint of the titration curve. The concentration of the semiquinone radical has a maximum at the midpoint and if we denote with k_s the semiquinone formation constant

$$k_s = \frac{[s]^2}{[r][t]} \quad (3)$$

the concentration of the radical at the midpoint will be given by

$$[s] = \frac{\sqrt{k_s}}{2 + \sqrt{k_s}} [a] \quad (4)$$

if $[a]$ is the concentration of the dye in all its forms.

Calculated values of the concentration of the semiquinone radical (for $[a] = 0.1$ molar), corresponding to different k_s values, are given in the following table. It is assumed that the dimerization is negligible.

$[a] = 0.1$ mole/liter

k_s	$\frac{\sqrt{k_s}}{2 + \sqrt{k_s}}$	[s] mole/liter	Expected conversion in mole liter ⁻¹ sec. ⁻¹
10	0.61	0.06	$6 \cdot 10^{-3}$
1	0.33	0.033	$3.3 \cdot 10^{-3}$
10^{-1}	0.14	0.014	$1.4 \cdot 10^{-4}$
10^{-2}	0.048	0.0048	$4.8 \cdot 10^{-4}$
10^{-3}	0.016	0.0016	$1.6 \cdot 10^{-4}$
10^{-4}	0.005	0.0005	$5 \cdot 10^{-5}$
10^{-5}	0.0016	0.00016	$1.6 \cdot 10^{-5}$

If we assume that the accuracy of determination at very slow rates is ± 10 per cent (in fact it is greater) the lower limit for the detection of the semiquinone radicals is about 10^{-4} mole/liter.

It will be seen that even with very small k_s the concentration of the semiquinone radicals is high enough to induce a well measurable conversion of para-hydrogen. Especially for k_s values lower than 0.01 is this of interest, since in this range the methods hitherto used cannot well distinguish between $k_s = 0$ and $k_s < 0.01$.

Some restriction to the applicability of the method must, however, be made. It is essential that the oxidation-reduction system should not undergo irreversible changes during the measurement of the conversion, which will take 1 to 24 hours. Furthermore it is obvious that only diamagnetic reducing (oxidizing) agents can be used for the oxidation (reduction) of the dye.



Restoration of the head of *Protoceratops andrewsi*. An old "male". One-eighth natural size. Georgia Mary Whitman, sculptress.

BROWN AND SCHLAIKJER: PROTOCERATOPS

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

VOLUME XL, ART. 3. PAGES 133-266

NOVEMBER 13, 1940

THE STRUCTURE AND RELATIONSHIPS OF
*PROTOCERATOPS**^{*}

BY

BARNUM BROWN AND ERICH MAREN SCHLAIKJER†

CONTENTS

	PAGE
INTRODUCTION.....	135
Discovery and Geological Occurrence of <i>Protoceratops</i>	135
Acknowledgments.....	138
Previous Publications on <i>Protoceratops</i>	138
Material Studied.....	139
COMPARATIVE STUDY OF THE SKULL ELEMENTS.....	139
Abbreviations of the Skull and Lower Jaw Elements.....	140
Rostral.....	140
Premaxillary.....	141
Maxillary.....	143
Lachrymal.....	147
Nasal.....	149
Prefrontal and Palpebral.....	154
Postorbital.....	156
Frontal.....	158
Parietal.....	166
Squamosal.....	169
Jugal.....	175
Epijugal.....	175
Quadratojugal.....	176
Quadrata.....	176
Exoccipital.....	177
Supraoccipital.....	181
Laterosphenoid.....	182

* Awarded an A. Cressy Morrison Prize in Natural Science in 1939 by the New York Academy of Sciences. Publication made possible through a grant from the income of the Nathaniel Lord Britton Fund, and the Centennial Fund.

† Assistant Professor of Geology and Paleontology, Brooklyn College, New York City.

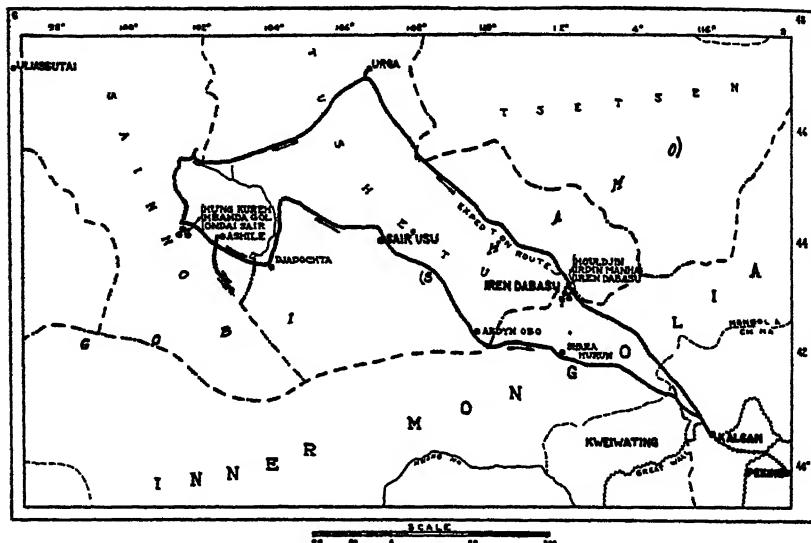


FIGURE 1 Route of the Third Asiatic Expedition in Mongolia, 1922, showing the Djadochta locality where *Protoceratops* was discovered and all subsequent material was collected. After Osborn.

Class Reptilia

Order Ornithischia

Protoceratops andrewsi Granger & Gregory

Pinacosaurus grangeri Gilmore

Order Saurischia

Velociraptor mongoliensis Osborn

Oviraptor philoceratops Osborn

Saurornithoides mongoliensis Osborn

Order Crocodilia

Shamosuchus djadochtaensis Mook

Order Chelonia

Dermatemyliid gen. indet.

Class Mammalia

Order Multituberculata

Djadocatherium matthevi Simpson

Order Insectivora

Deltaitheridium pretrituberculare Gregory & Simpson

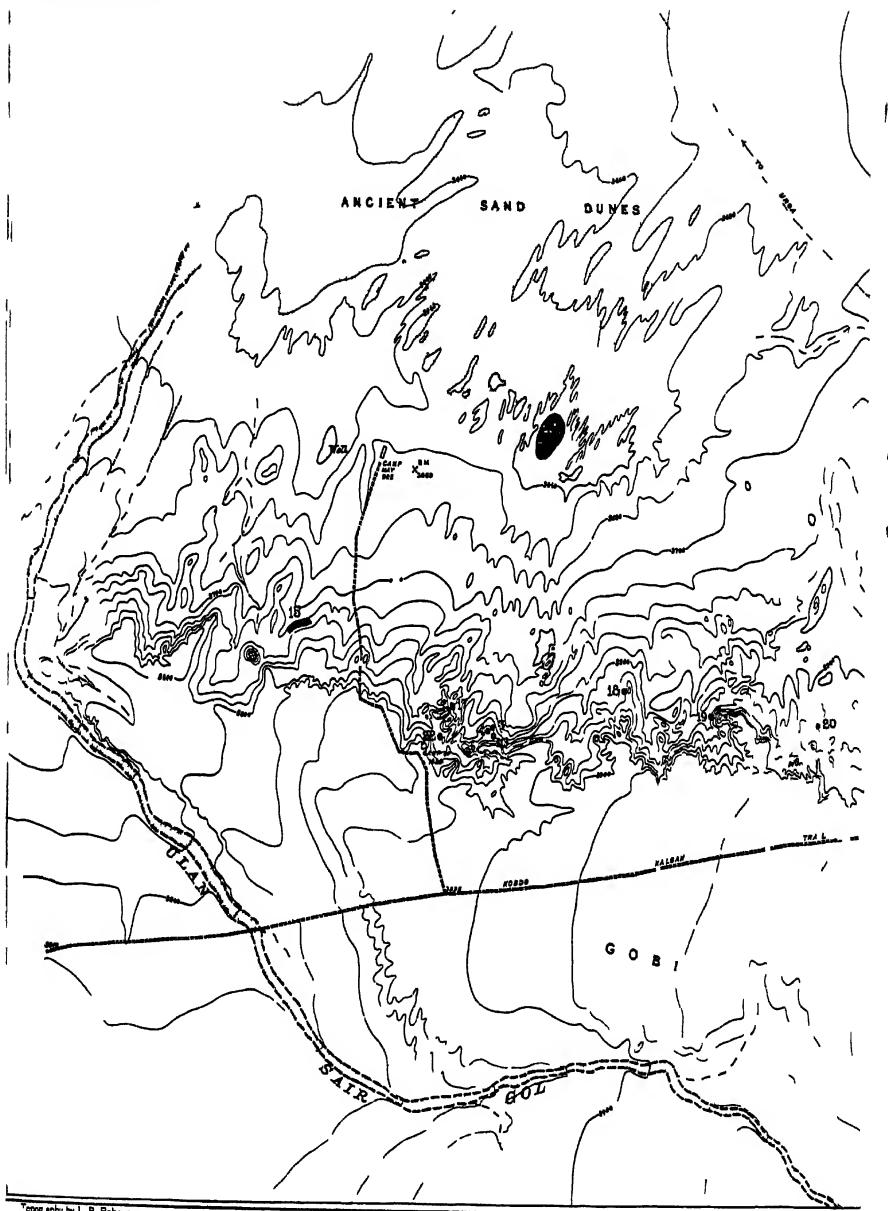
Deltaitheridium cretacicus Gregory & Simpson

Hyotheridium dobsoni Gregory & Simpson

Zalambdalestes lechei Gregory & Simpson

Zalambdalestes grangeri Simpson

The Djadochta beds are approximately 500 feet in thickness and are composed mostly of red sandstone. Concerning the lithology and



Topography by L. B. Robe
by F. B. Battle and H. O. Robinson

Contour

FIGURE 2 Detailed map of the Djadochta locality

MONGOLIA

1925

MAP OF
DJADOKATA FOSSIL FIELD
AND
SHABARAKH USU ARTIFACT
AREA

TYPICAL UPLAND AND ESCARPMENT TOPOGRAPHY
WITH LOCAL BADLANDS DEVELOPMENT AND
A SAND OBSTRUCTED VALLEY

U P L A N D

No	AM MUS No
3	6638
4	6508
10	6418
11	6467
12	6471
13	6466
14	6416
15	6427, 6822 6513, 6638 6639, 6640
17	2905
18	6417
19	6408
20	6414
21	6631
22	6251

Elevations are based on Pek-n-Suyan railway levels
To refer to mean sea level add 31 feet
Starting a from memory banner at Kalgan



gives locations of the principal specimens collected

probable origin of this formation, Berkey and Morris make the following statement (1927: 157-158):

"The rock is a red sandstone of very uniform grain and comparatively simple structure. The grains are exquisitely graded, so that when the stone is weathered, it yields a sand that will run like the sand in an hour glass. There is virtually no admixed clay, and separate beds of clay are few, occurring chiefly as channel fillings. Many of the sandstone beds contain myriads of small limy concretions, some of them being traceable for as much as a mile. In places the sandstones appear massive and structureless, forming sheer vertical walls as much as forty feet high, in which neither bedding nor color-banding is seen; yet there massive layers pass laterally into well-bedded deposits. Cross-bedding, apparently of aeolian type, is developed on a large scale at certain horizons. We believe that the formation is in large part wind-blown, and that this history is the major factor in accomplishing the perfect preservation of the delicate fossils which the Expedition recovered from it."

The exact age of the Djadochta formation is questionable. A survey of the list of the species given above, however, favors unmistakably an assignment to the Upper Cretaceous. Among the twelve species recorded, only *Protoceratops andrewsi* gives any hint at all of possibly more definite correlation. As will be shown later, *Leptoceratops* from the Edmonton of Alberta, is indeed a very close relative of this Mongolian species. This affinity, therefore, suggests that the Djadochta beds are of Edmonton age. If, on the other hand, Mongolia is regarded as the center of dispersal of the horned dinosaurs—which is quite probable—then it is to be expected that the beds containing the archaic *Protoceratops* would be somewhat older than those in which the North American slightly more progressive *Leptoceratops* is found. Providing, of course, that *Protoceratops* is not a primitive survivor from an earlier period when such dispersal might have taken place. Offsetting this suggestion, however, is the possibility that *Leptoceratops* itself was a living fossil in the Edmonton from an earlier period of emigration to North America. In this case, the Djadochta formation would be considerably older than the Edmonton.

The large and splendidly preserved *Protoceratops* collection is indeed unique. Every important growth stage from the egg to the old individual is represented by numerous specimens. No other collection of a fossil reptile is so complete. Moreover, this collection is even more remarkable because all of the material was secured in one locality. The specimens were scattered along an escarpment no greater than five miles in extent, and occurred within a maximum vertical range of only 150 feet in a rock of decidedly uniform lithology. These facts corroborate our conclusion, from a study of the material, that no more than one species is represented.

Acknowledgments

We are especially indebted to Dr. Roy Chapman Andrews, leader of the American Museum Asiatic Expeditions, and to Dr. Walter Granger, paleontologist of those Expeditions, for the privilege of studying the *Protoceratops* collection. Nearly every one in the paleontological laboratory of the Museum has had some part in the preparation of this material, most of which, however, was done by the late Peter C. Kaisen. Mr. Charles J. Lang is especially responsible for the preparation and mounting of the skeletons, and Mr. Otto Falkenbach has been particularly valuable in revealing the sutures on many specimens, and in skilfully disarticulating the cranium of one of the larger skulls and obtaining from it a most complete endocranial cast. Mr. Alastair Brown is solely responsible for the numerous splendid drawings in this paper. In many instances before illustration of certain characters was possible, considerable detailed preparation of specimens was required, which he also meticulously executed. Sculptress Georgia Mary Whitman has successfully produced the life-like heads showing three growth stages of *Protoceratops*.

Previous Publications on *Protoceratops*

Numerous references have been made to *Protoceratops* in both scientific and popular writings. Only three brief papers have appeared, however, which are especially devoted to a study of this primitive ceratopsian. The first of these is a preliminary description of the type by Gregory and Granger (1923) containing a short account of the structural relations of the Djadouchta beds by Berkey. The second brief paper, by Gregory and Mook (1925), is devoted to the designation of the family Protoceratopsidae, to a synopsis of outstanding characteristics of the genus, and to a summary of general ceratopsian relationships. The third paper consists of a four-page description of the microstructure of the egg-shells by Van Straelen (1925). In addition to these, Gregory gave considerable space to the morphology and evolutionary position of *Protoceratops* in his paper on "The Mongolian Life Record" (1927). Lull, of course, in his monograph on the Ceratopsia (1933) devoted a short section to *Protoceratops*. This consists mainly of a slight revision of the generic characters, and quotations from Gregory and Mook with a few additional observations. Two papers by Brown and Schlaikjer have certain sections in which *Protoceratops* is specifically treated, and should be mentioned here. They are, "The Origin of Ceratopsian Horn Cores" (1940a), and "A New Element in the Ceratopsian Jaw with Additional Notes on



Djadochta formation exposed in the Flaming Cliffs at Shribnakh Usu

BROWN AND SCHLAIKER *PROTOCERATOPS*



Cliffs at Djadochta showing the massive, fine, red sandstones of the Djadochta formation
Photograph by Walter Granger

BROWN AND SCHLAIKER *PROTOCERATOPS*

the Mandible" (1940b). A semi-popular article by Granger (1936) is devoted to "The Story of the Dinosaur Egg".

Material Studied

Our detailed study of *Protoceratops* has been supplemented by a thorough re-examination of all the ceratopsian material in the American Museum, and by reference to the more important specimens in the National Museum, Peabody Museum of Natural History, Yale University, Royal Ontario Museum of Paleontology, and Museum of Comparative Zoology, Harvard University.

A complete list of all the *Protoceratops* specimens is given in the appendix.

COMPARATIVE STUDY OF THE SKULL ELEMENTS

While later in the paper summaries of skull characters are given, it is well to point out here that in the following pages the procedure has been to describe each element of the skull from the young to the adult stage, emphasizing especially the salient growth changes. The primitive and variable characters are cited, and in each case an attempt has been made to point out the ontogenetic characters wherein *Protoceratops* foreshadows the evolution of the later ceratopsians. Stress is also given to the importance of certain elements in the architecture of the skull, and in each of its many characteristics, *Protoceratops* is compared with the later ceratopsians.

Sutures are clearly discernible in nearly all of the skulls. It is of interest to note, however, that in an occasional young skull certain sutures are closed, while in some adult skulls these sutures remain open. Whether sutures are closed or open, is not, therefore, a safe criterion to age determination.

The determination of skulls as male or female is purely a matter of conjecture, and our reasons for so doing are indeed tenuous. We regard those skulls as males which at any growth stage—as judged purely on a basis of size—are most robust and which give greatest emphasis of the salient growth changes such as, widening of the frill, development of a parieto-frontal depression, increase of facial depth, development of incipient nasal horn-cores, etc. This conclusion is in keeping with the tendency among some reptiles, that the males are larger and more robust than females. It should be remembered, however, that in certain living forms the males are smaller and more delicate. This might also be true for *Protoceratops*. Another interpretation is that this supposed sexual distinction is nothing more than

individual variation. This, however, seems less likely since each of the types presents some rather marked variations.*

The unique and bizarre ceratopsian skull has presented paleontologists with many morphological enigmas. The large assemblage of splendidly preserved *Protoceratops* skulls has indeed afforded much information relevant to these riddles. In the present paper, particular emphasis has been given the following three more important problems, origin of horn-cores, composition of the frill, and origin of the secondary skull roof. These are treated in the discussion of the particular element or elements involved. The problem of the origin of horn-cores has been more fully considered in another paper (Brown & Schlaikjer 1940a).

From this study it has become evident that the skull of *Protoceratops*, as suggested by Professor W. K. Gregory (1927: 177-180), in all of its forty-six elements is a remarkably ideal prototype for the Ceratopsia.

Abbreviations of the Skull and Lower Jaw Elements

Rostral.....	<i>r</i>	Laterosphenoid.....	<i>ls</i>
Premaxillary.....	<i>pmx</i>	Prootic.....	<i>pc</i>
Maxillary.....	<i>mx</i>	Basioccipital.....	<i>bo</i>
Lachrymal.....	<i>lac</i>	Basisphenoid.....	<i>bs</i>
Nasal.....	<i>nas</i>	Pterygoid.....	<i>pt</i>
Prefrontal.....	<i>prf</i>	Ectopterygoid.....	<i>ec</i>
Palpebral.....	<i>pap</i>	Palatine.....	<i>pl</i>
Postorbital.....	<i>po</i>	Prevomer.....	<i>pvo</i>
Frontal.....	<i>f</i>	Predentary.....	<i>pd</i>
Parietal.....	<i>pa</i>	Dentary.....	<i>d</i>
Squamosal.....	<i>sq</i>	Angular.....	<i>ang</i>
Jugal.....	<i>j</i>	Surangular.....	<i>sang</i>
Epijugal.....	<i>ej</i>	Articular.....	<i>art</i>
Quadratojugal.....	<i>qj</i>	Prearticicular.....	<i>part</i>
Quadrata.....	<i>q</i>	Splenial.....	<i>sp</i>
Exooccipital.....	<i>exo</i>	Coronoid.....	<i>cor</i>
Supraoccipital.....	<i>so</i>	Intercoronoid.....	<i>icor</i>

Rostral

The rostral is considerably rounded in front so that the inferior tip of the beak lies almost below the thin dorsal point. This dorsal point covers over, and beneath is wedged between, the anterior dorsal projections of the nasals. The dorsal tip extends upward to a point slightly below the middle of the anterior margin of the narial opening.

* The fact that the two types of skulls are approximately equally represented in numbers also suggests the sex ratio 1 : 1.

It approaches close to, but is never in contact with, the anterior tips of the nasals. The postero-inferior projections diverge outwardly at a rather sharp angle and extend back to the first of the two alveoli in each of the premaxillaries, which brings them into close proximity with the antero-inferior ends of the maxillaries. The lateral wall covers a considerable part of the anterior end of the premaxillary which is received onto a low shelf-like ridge inside. There is a short blunt median inferior projection which wedges between the premaxillaries. The posterior margin is gently curved anteriorly. The external surface is quite rugose.

In the later ceratopsians there is a tendency for the rostral to occupy a more inferior position on the skull. This is a feature that is in keeping with the enlargement and expansion of the anterior portion of the premaxillaries and an enlargement of the nasal openings. Likewise, as a result of these changes, the rostral in the later forms becomes greatly separated from the nasals above, and from the maxillaries below. Also, in the later forms the rostral becomes more rugose, the anterior end relatively more robust, and the posterior margin more sharply curved anteriorly.

Premaxillary

In *Protoceratops* the premaxillary occupies a greater proportion of the face than that bone does in any other ceratopsian. The posterior branch of the facial portion is broader than the anterior branch, and it extends upward and backward to a line even with or above the dorsal margin of the lachrymal, a feature, insofar as is known, that is distinctive of this genus. Although in some cases, especially in the skulls of the younger individuals, it is closely approximated to, it is never actually in contact with the lachrymal. The only recorded instance of such a contact is in *Monoclonius* (*Centrosaurus*), and, as shown by Lull (1933: 32), it is a variable feature of that genus.

The anterior branch extends upward and abruptly backward to a position a short distance in front of the posterior margin of the narial opening in the small skulls and grows beyond the posterior margin of the narial opening in the older individuals. It is tightly and extensively wedged in between the anterior ends of the nasal bones. The width of this branch increases from the young to the adult stage. This results primarily from a change in position of the oval-shaped narial opening, for the long axis of this opening becomes more erect with age. As this takes place, the anterior branch of the premaxillary increases in width, and there is a corresponding decrease in width of the pos-

terior branch. In what we regard as the male skulls, the long axis of the narial opening is proportionately more erect than in the female skulls. The probable explanation of this is that the face of the male skull is proportionately shorter and deeper.

Immediately below and in front of the antero-inferior margin of the narial opening there is a shallowly depressed area. This seems to be the homologue of the fossa in front of the narial opening which is so extensively developed in the later ceratopsians. In these, from a primitive form such as *Brachyceratops* to the advanced *Triceratops*, the narial opening migrates downward and backward as the face deepens, and as the nasals enlarge in the formation and support of the nasal horn. By this change, the premaxillary becomes profoundly modified. The posterior wing is crowded downward, and the fossa becomes deepened, greatly enlarged, and confluent with the narial opening. The anterior parts of the premaxillaries thus become greatly enlarged, and medially in the fossa they become thinly compressed posteriorly to form a septum. In *Protoceratops* there is no septum, although there is a beginning of one, in the form of a welt, inside the narial opening, at the antero-inferior end where the premaxillaries meet. As the fossa deepens in the later forms and becomes merged with the narial opening, the premaxillary forms a thin septum which at first, in forms such as *Brachyceratops* and *Monoclonius*, is confined to a position in front of the narial opening, but later grows back to form a postero-inferior wing in the narial opening. This structure is primitively shown in *Chasmosaurus* and *Pentaceratops*, is more advanced in *Arrhinoceratops*, and is excessively developed in *Triceratops*. In some of the earlier divergent genera and in the progressive genus *Triceratops*, the septum in the fossa is perforated by an intermaxillary fontanelle. Concerning this, Lull (1933: 32) says, "This fenestration is not present in all *Triceratops* skulls, which leads to the supposition that it may not have been normally present in the living animal, but is merely an accidental post-mortem perforation through the thin septum." As admitted in part of this statement, and as is unquestionably shown in some of the specimens, this perforation is a true fenestra in some skulls. Its absence in others simply signifies that it is a variable feature and is, in some individuals, merely a further expression of marked fenestration in the whole front part of the skull—a requisite for strengthening in that region which was in keeping with increase in size of brow horns, and increase in masticatory function of the jaws.

The inferior margin of the premaxillary is short in *Protoceratops*. In the older individuals it is straight, but in the younger ones it is

occasionally convex downward, although it is normally concave upward. The latter is principally the result of the development of a rather sharp projection at the posterior margin at the union with the maxillary. A feature of the ventral margin of the premaxillary in *Protoceratops* that is unique among the ceratopsians is the presence of two subequal, rather long, conical teeth that are set close together midway between the posterior end of the rostral and the front of the maxillary.*

On their ventral surfaces, the premaxillaries unite to form the narrow and deep front portion of the skull. This narrowness and deepness is more emphasized in the skulls of younger individuals. Postero-medially they are separated by a projection of the maxillaries that extends forward to opposite the anterior of the front premaxillary tooth. This is a primitive feature, for in later forms the projections of the maxillaries become much shortened and more blunt. Also, the rounded, and backwardly curved posterior margin of the premaxillary becomes fairly straight in *Monoclonius*, and in *Triceratops* it actually curves forward. On the palatal surface of the premaxillary there is a single large foramen opposite and a little in front of the posterior projection of the rostral. In *Triceratops* there are two and sometimes three such foramina.

Maxillary

The maxillary is proportionately deeper than in any other known ceratopsian. It extends four-fifths up the side of the face, at least to opposite the middle of the front margin of the orbit where it has a slight contact with the nasal—a contact that seems to be of no particular taxonomic importance in the Ceratopsidae. In some of the later forms this contact is prevented by premaxillary-lachrymal union, and in others it is not.

Steepness is characteristic of the front of the maxillary. In later genera, along with the enlargement of the narial opening and downward crowding of the posterior wing of the premaxillary, the front margin obtains a very gentle slope. This character is most emphasized in *Triceratops* and especially in the species *T. serratus*. Steepness is likewise a feature of the posterior margin but, again, this is primitive, for in contrast to later forms, the posterior alveolar portion is not as extended, and the jugal is not erect and not as constricted at the point of contact with the maxillary.

* Fragmentary material in The American Museum collection and in the National Museum shows that premaxillary teeth were also present in *Leptoceratops*.

From the ventral margin of the jugal a heavy ridge extends forward onto the maxillary. It becomes less pronounced anteriorly, and under the front of the antorbital fossa it bends downward and reaches the inferior edge of the maxillary just in front of the teeth. The development of this ridge on either side of the face gives the whole front part of the skull a wedge-shaped appearance. In his restoration of *Diceratops hatcheri*, Lull (1905: 422) intimated that these and corresponding ridges on the dentaries showed the presence of "muscular cheeks" in the Ceratopsia. In his study of cranial musculature in the ceratopsians (1908: 389) he states, "Cheek muscles must have existed, originating on the outer side of the maxillary and the posterior portion of the premaxillary as indicated by a sudden inward compression of the lower portion of these bones along a line running obliquely downward and forward from the jugal to the lower margin of the premaxillary bone. The insertion of this muscle lies along the forward margin of the coronoid and sweeps forward along the outer surface of the jaw, finally rising again to the end at the upper termination of the dentary-predentary suture. This broad sheet of muscle, probably equivalent to the buccinator, was subsidiary to mastication, as its chief function was to retain the food in the mouth. The extent of this muscle limits the backward extent of the gape of the mouth, as the writer (Lull 1905) has shown in previous papers." On plate one in this paper, the supposed muscle is definitely labeled "buccinator." Though with no occasion to refer to this muscle by name, Lull, in his "Revision of the Ceratopsia or Horned Dinosaurs" (1933: 21), retains his belief in "muscular cheeks," which are "formed largely of the masseter muscles."

The most recent contribution to the subject of musculature in the ceratopsians is by Russell (1935) who describes (p. 40) "A sheet-like muscle, which extends vertically from maxilla to dentary, with a possible attachment to the angle of the mouth." This he definitely calls, and illustrates as (figure 9), the buccinator muscle.

As is well known, the buccinator muscle is not present in reptiles. It is a facial muscle, innervated by a branch of the seventh nerve, that is characteristic of the Mammalia. It is a member of the group of facial muscles derived from a muscle sheet, known as the sphincter colli, situated on the side of the neck in reptiles.*

If a muscle were present on the side of the face of ceratopsians, it

* Among the many contributions written on the evolution of head and neck musculature, two outstanding and thorough works are: "A Memoir on the Phylogeny of the Jaw Muscles in Recent and Fossil Vertebrates," Adams (1919), and "Evolution of Facial Musculature and Cutaneous Field of Trigeminus," Huber (1930).

would have to be a branch of the superficial layer of the capiti-mandibularis. This is possible, but highly improbable. The presence of such a muscle as a separate vertical sheet, as restored by Lull and Russell, is certainly improbable. Since the width across the maxillary ridges is greater than that across the ridges leading forward from the ascending portions of the dentaries, muscles from the maxillary ridges to the ridges on the dentaries would have had a tendency to pull the jaws outward and upward. We know the capiti-mandibularis to have been very well developed. It exerted an upward and backward pull on the jaw while the large pterygoideus anterior extended obliquely across this and exerted an upward and forward pull. They thus cooperated in closing the jaw. A maxillary-dentary muscle would have usurped much of the upward pull of these two larger muscles, and would have pulled outwardly against them both—a most unnatural function. Furthermore, there is no indication of a muscle attachment on the maxillary. Just above the alveoli, extending back about halfway along the tooth-row, there are a number of foramina for the emission of small branches from the maxillaris branch of the trigeminus nerve—showing that the labial skin was highly innervated. Similar foramina (for emission of branches from the mandibularis branch of the trigeminus) are present on the dentary back to about the middle of the tooth-row. Behind this area the coronoid ridge ascends rapidly, and along the inside of it and in the coronoid area is a well-marked zone of insertion. We believe that this marks the insertion of the capiti-mandibularis which extended upward and backward and occupied the large trough that leads under the jugal, covering the posterior part of the alveolar portion of the maxillary, and on up to the parietal crest. We are also of the opinion, for reasons cited above, that there is no evidence that a muscle from the maxillary ridge to the ridge in front of the ascending portion of the dentary was present in any of the ceratopsians. The maxillary ridges seem to be nothing more than braces in front of the heavy inferior margin of the jugals—normal structures in the architecture of such a skull.

The inferior margin of the maxillary curves outward and forward in front of the first alveolus for a distance of approximately two-sevenths of the greatest length of the maxillary. In later ceratopsians this pre-alveolar space is much reduced, since the first tooth is situated almost at the premaxillary-maxillary suture.

On the ventral surface the maxillaries unite medially to form an anterior extension wedged between the premaxillaries. The form and extent of this projection is somewhat variable. It is V-shaped and

fairly blunt. Normally, it extends forward to opposite the first premaxillary tooth (a primitive feature). The palatal contact of the maxillaries is restricted to only the median projection. Behind this the paired prevomers are wedged tightly between the two bones back to the opening of the internal nares which begin about opposite the first maxillary tooth. In *Chasmosaurus belli* (Lull 1933, fig. 30) the maxillaries have only a tip of the prevomers wedged between them, and in *Triceratops serratus* (Am. Mus. No. 970) none on the palatal surface.

In *Protoceratops* the maxillaries flare outward posteriorly giving the palate an open V-shaped form. With age the palate becomes proportionately even wider posteriorly—a change that is correlated with the proportionate increase in width of the whole skull in the adult specimens. In the later ceratopsians the position of the maxillaries is more antero-posterior, and the posterior part of the palate is relatively much narrower. This is the result mainly of the development of a narrower skull and the posterior growth of the maxillaries.

On the palatal surface just above the alveolar border there is a series of "foramina"—one for each alveolus. These are irregular and small, although in most specimens they appear to be large because the very thin margins have been broken away. These are undoubtedly for the emission of branches of the maxillary branch of the trigeminus. Their serial arrangement, their larger size under larger teeth, and their change from a rounded form in the young to an oval form in the adult suggest that they originated from the dissolving of the bone at the base of each vertical series of teeth when the animal's growth rate was most rapid. (See Brown and Schlaikjer 1940b for a full discussion of analogous foramina in the lower jaw.)

Ventrally the posterior surface of the maxillary is mostly in contact with the palatine. Below this, in the immature specimens, it is braced against the ectopterygoid. With age there is an increase in number of teeth and a proportionate increase in the alveolar dimension of the maxillary. It grows posteriorly and on the extended alveolar margin becomes braced against the anterior latero-ventral projection of the pterygoid, thus restricting the palatal extent of the ectopterygoid. This is precisely what happens in the more advanced ceratopsians, only to a greater degree. In *Styracosaurus* (Lambe 1913: 112) and in *Triceratops* (Hatcher, Marsh & Lull 1907: 26), the ectopterygoid is eliminated entirely from the palatal surface and the maxillary is more extensively in contact with the pterygoid.

The greater portion of the large preorbital fossa is developed on the

maxillary. It also includes a considerable portion of the lachrymal, and the most anterior external point of the jugal. Inferiorly it is deep and pocket-like and is bounded by a strong shelf-like border just above the maxillary ridge. A similar border overhangs the superior margin, thus forming a rather deep pocket. In the older individuals there is a tendency for this pocket to become eliminated and for the margin to become less distinctive. This seems to foreshadow what takes place in the later ceratopsians, in which the fossa is almost entirely eliminated, with only the preorbital foramen remaining. The foramen is situated posteriorly in the fossa where the lachrymal, maxillary, and jugal sutures meet. It varies slightly in size and position, and leads directly into the narial chamber. In the younger forms the foramen is relatively larger and the fossa is deeper and better defined. These facts seem to substantiate the suggestion of Gregory (1920: 127) that the fossa is a vestige of an antorbital opening.

Lachrymal

The relatively large lachrymal is another primitive feature of *Protoceratops*. It occupies most of the front margin of the orbit and a considerable part of the face. It is somewhat proportionately smaller in the older individuals—a tendency that is prevalent among the later ceratopsians. Also, in these later forms the orbital extent of the lachrymal becomes limited, the ventral margin straightened, and the dorsal margin, which is quite straight and meets the anterior margin about at a right angle in *Protoceratops*, becomes steeply sloping and unites with the anterior margin to form a wide open angle. These changes take place as the anterior portion of the jugal swings forward, as the preorbital fossa is eliminated, and as the nasal enlarges and grows downward onto the side of the face. (See FIGURE 3.)

In the ceratopsians the lachrymal acts as a sort of keystone in the architecture of the orbit, and undergoes a considerable modification throughout the evolution of the group. In *Protoceratops* it is quadrangular in outline, and in the very young individuals the front border faces obliquely downward. With age, the front border becomes more erect until in the very old forms it is nearly vertical. This change of the lachrymal to a more erect position seems correlated with deepening of the face, and the enlargement, especially the upgrowth, of the nasals. In skull Am. Mus. No. 6429, though a fairly mature individual, the face is not as deepened as in other specimens of the same size and the front of the lachrymal faces much more obliquely downward as in the young forms. In all members of the group the lachrymal receives

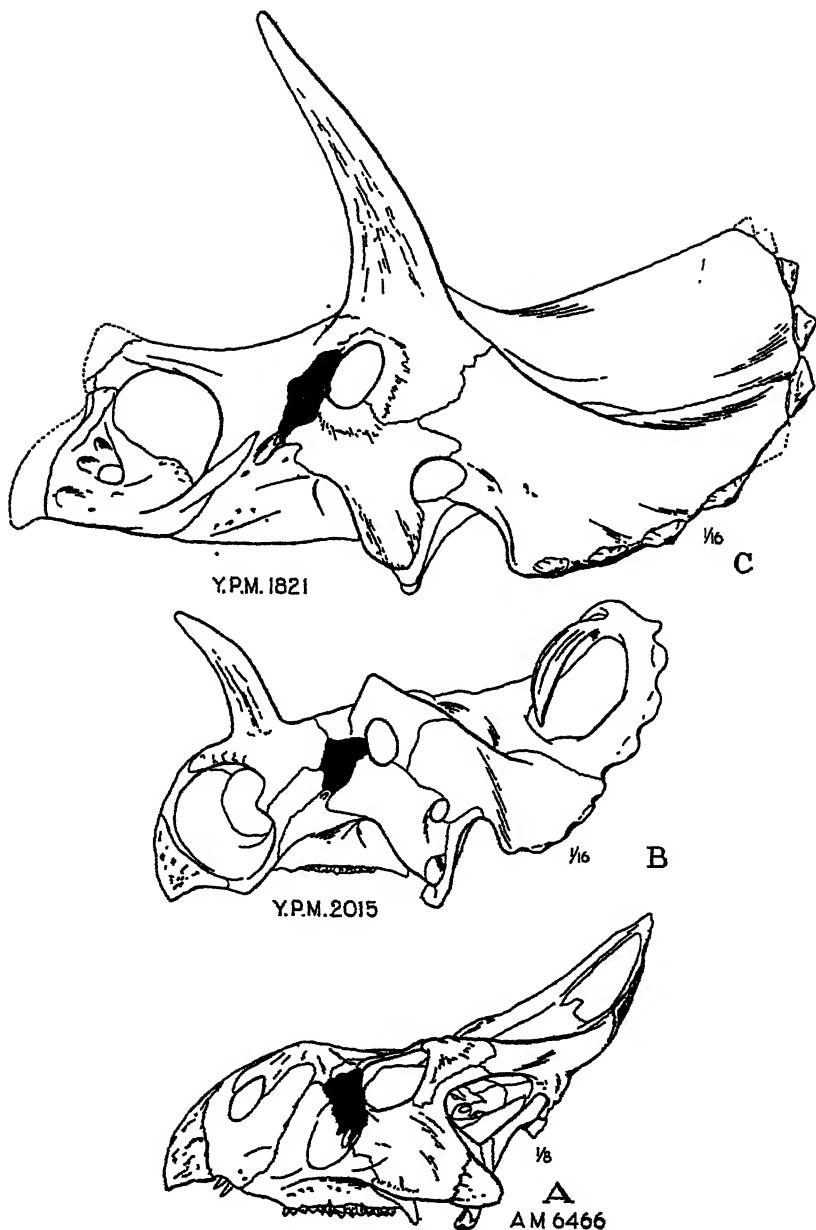


FIGURE 3. Series of ceratopsian skulls showing the reduction and change of position of the lachrymal bone. A, *Protoceratops andrewsi*. B, *Monoclonius flexus*, modified from Lull. C, *Triceratops flabellatus*, modified from Hatcher, Marsh, and Lull.

stresses from the jaw articulation through the large jugal, from the maxillary below, and from the temporal area via the supraorbital arch. The development of horns and their related skull modifications add to the functions of the lachrymal, and reflect in it corresponding changes. In the older individuals of *Protoceratops* there was at least an incipient horn-protuberance covering the highly arched nasals. From this area stresses were transmitted, partially via the prefrontal, through the lachrymal backward to the jugal, and downward to the maxillary. In a more advanced form, such as *Monoclonius*, the front of the lachrymal begins to face obliquely upward, and the entire bone begins to lose its quadrangular shape. The anterior and superior margins meet to form an obtuse angle wedged between the latero-posterior margin of the enlarged nasal and the latero-anterior margin of the frontal—a strategic position for transmitting stresses from the large nasal horn postero-ventrally to the jugal. Stresses from the rudimentary brow horn above can also be relayed to the lower bones of the face. In the advanced *Triceratops*, of the uppermost Cretaceous, in which the nasal horn is still present, and the brow horns are enormously developed, the lachrymal is quite lozenge-shaped. With the forward rotation of the jugal and the downward progression of the premaxillary and nasal, it is admirably located for sending to the maxillary and jugal stresses received from the brow horns, and for disseminating around the orbit those received from the nasal horn.

Nasal

As in *Leptoceratops* the nasal is smaller than in any of the other known ceratopsians. In lateral view, it is quite short and fairly deep in the very young individual. In the medium-sized skulls it lengthens and is relatively less deep. In the older skulls it begins to arch up about midway back forming an incipient horn-core, and assumes deeper proportions. As this takes place, the anterior extension becomes somewhat abbreviated and an antero-inferior process grows down below the posterior part of the narial opening, causing the lateral form of the nasal to become forked in front. Likewise, there are marked changes in proportions in the dorsal form of the nasals from the young to the adult stage. In the young specimens, the nasals are broad, flat, and bluntly wedge-shaped. With age they become proportionately narrower and longer, although this change is not especially constant with age. Occasionally a fair-sized skull, such as Am. Mus. No. 6429, has the short and wide nasals more as in the juvenile forms. This shows that there is some variation in the nasal, which

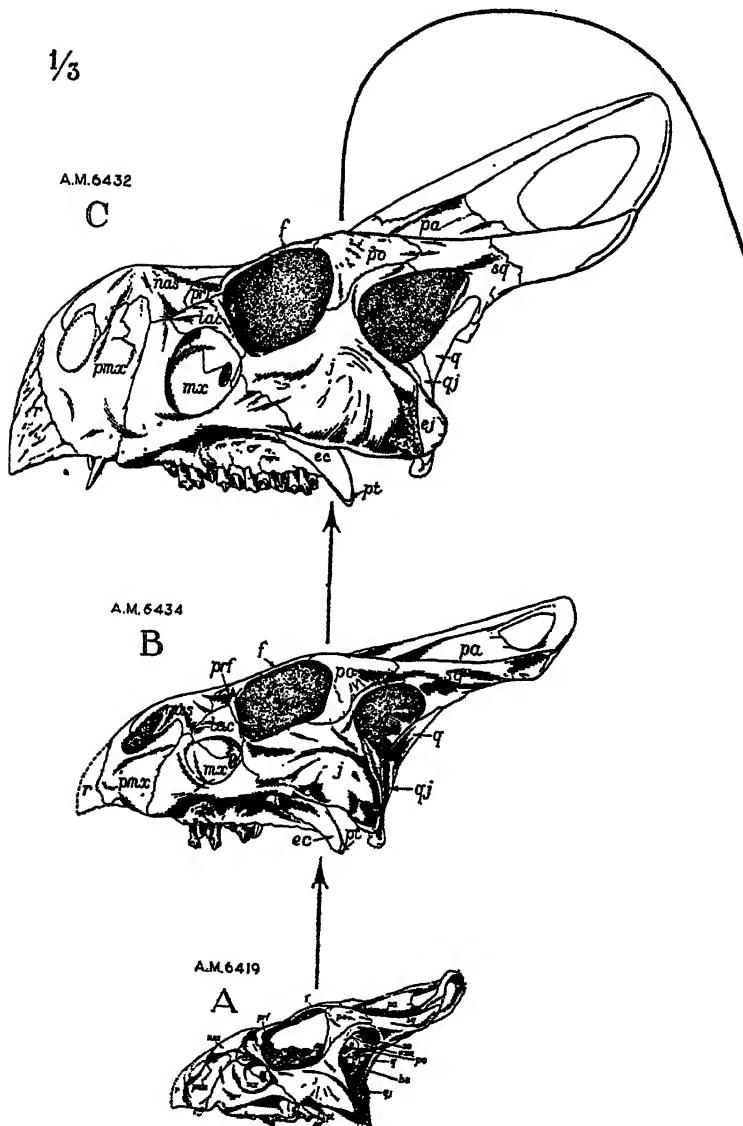
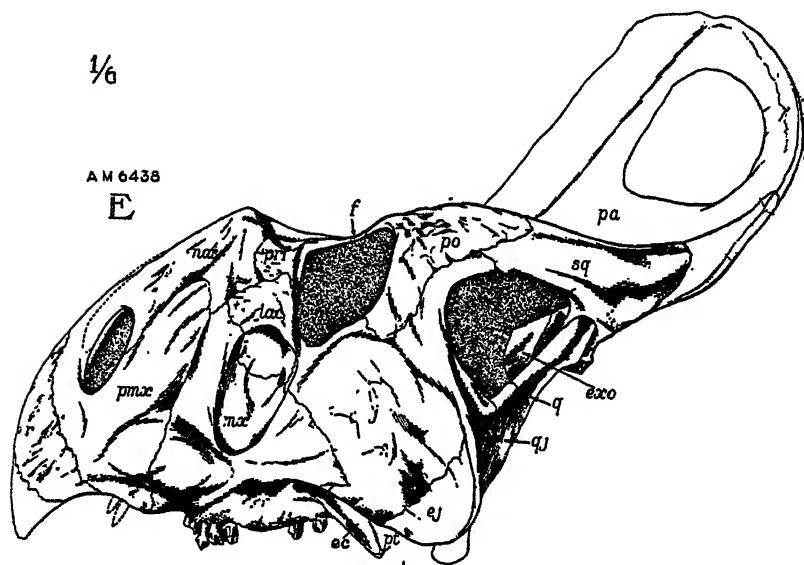


FIGURE 4. Series of *Protoceratops* "male" skulls showing the development from a very immature (A) to an old (C) individual.



A.M.6466

D

simply means that the tendency for that bone to become narrower and longer is somewhat retarded in an occasional skull.

Anteriorly the nasals embrace the posterior extensions of the premaxillaries, and the extent of this embracement increases with age so that in the skull of an old individual it is considerably behind the posterior margin of the narial opening. Posteriorly the frontals are bluntly wedged in between the nasals to a point just anterior to opposite the front of the orbit in the very immature skulls. As the nasals elongate, this point shifts posteriorly to opposite or behind the front border of the orbit. There is a backward migration of the posterior projection of the nasal which also becomes more pointed with

There is a longitudinal median groove on the dorsal surface of the nasals that extends throughout their sutural length. In the young and early adult skulls, this groove continues back onto the anterior projection of the frontals. In the older individuals, it is confined to the nasals. In the smallest skulls, this groove is broad and very shallow. With age it becomes deeper and narrower, although there is some variability in deepness and narrowness in both, what we consider as male and female skulls. In skulls of the same size, however, it does seem to be deeper and narrower in the male. The development of this groove, together with the proportionate deepening and narrowing of the nasals and the marked upward arching of them into an incipient horn-core, is of considerable morphological significance insofar as the evolution of the nasal horn in the ceratopsians is concerned. As the nasal bones arch upward, the median groove becomes very constricted between the apices of the convexities. This arching of the nasals and the fact that the grain of the bone tends towards the apices suggests that the nasals probably bore one, or possibly two horn-like protuberances. Although there is no evidence that ossicles were present.

In the light of this evidence, it seems reasonable to conclude therefore, that the arching of the nasals in *Protoceratops* represents the beginning of a nasal horn-core in the ceratopsians. The only change necessary for giving rise to the next advanced stage, as seen in *Brachyceratops*, would be for the nasals to continue to grow upward into a pronounced laterally flattened nasal horn-core. (See FIGURE 5.)

The problem of the origin and evolution of the ceratopsian nasal horn-core has been more fully treated in a previous paper (Brown and Schlaikjer 1940a).

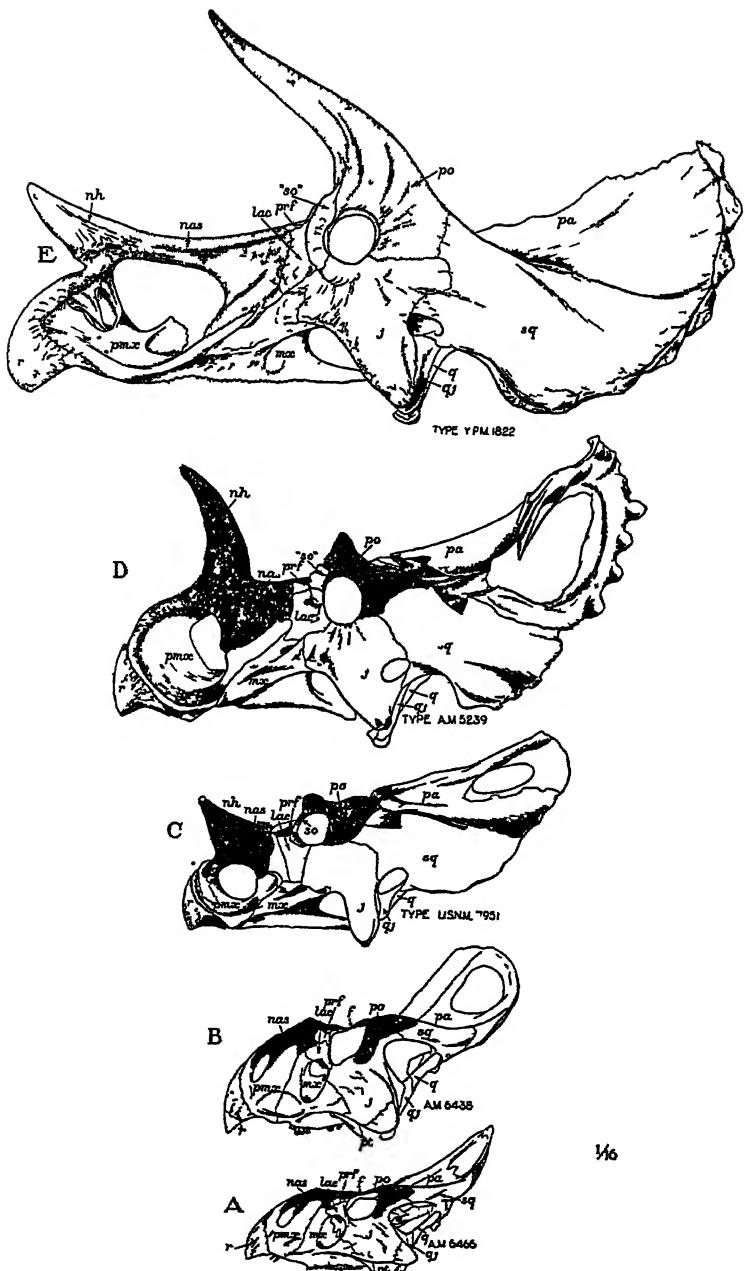


FIGURE 5. Structural series of ceratopid skulls showing the development of horn-cores. A, *Protoceratops andrewsi*. Skull of an adult "male." B, *Protoceratops andrewsi*. Skull of an old "male." C, *Brachyceratops montanensis*. Modified from Gilmore. D, *Monoclonius flexus*. Modified from Lull. E, *Triceratops prorsus*. Modified from Hatcher, Marsh, and Lull. From Brown and Schlaikjer 1940a.

Prefrontal and Palpebral

The prefrontal shows marked changes of proportions with age. In the skull of a young individual it is relatively short, proportionately broad in front, extends down on the side of the face somewhat, and on the orbital border it is restricted more to the antero-superior corner. With age, it elongates posteriorly, thus approaching the postorbital more closely; it becomes narrower in front; and, in a fully adult skull, it forms more than one-half of the superior border of the orbit. This change in form and position is controlled by a proportionate reduction in the size of the orbit, the elongation of the nasal posteriorly, and the backward shifting of the frontal.

The prefrontals do not meet in the midline to exclude the nasals from contact with the frontals, although they more closely approximate the midline with age. This character is unique in *Protoceratops*.* In all known later forms the prefrontals unite along the midline on the skull surface between the nasals and frontals. In *Brachyceratops* the union is just beginning, but in the more advanced forms it is extensive. In such a later stage the prefrontals act as braces across the top of the skull just in front of the much enlarged brow horn-cores thus meeting the demand for greater strengthening in this area. This change proceeds with the modifications, to be described below, which the postorbitals and frontals undergo during the evolution of the Ceratopsia.

The fact that the prefrontal forms such an extensive part of the orbital margin in *Protoceratops* is also unique. In the known later genera, according to von Huene, Gilmore, Granger and Gregory, and Sternberg, the antero-superior margin of the orbit is formed by a separate element, called the "supraorbital," which separates completely the prefrontal from the margin of the orbit. Von Huene (1911) figures (figure 3) a supraorbital in the skull of *Triceratops prorsus* (Y.P.M. 1820), and says (p. 159) that one is present in the skull of *T. horridus* (Y.P.M. 1820), and in a fragmentary specimen of *Triceratops* sp. (U.S.N.M. 4286). In describing the prefrontal of *Brachyceratops*, Gilmore (1917: 9-10) says, "Near the posterior termination on the external side a narrow vertical sutural surface was for the articulation of the small supraorbital bone, which is missing. This element would have completed the thickened orbital border which projects immediately in front of the eye and which forms so conspicuous a feature of

* The prefrontals of *Leptoceratops* are unknown. The sutural surfaces for contact with them that are preserved on the nasals seem to indicate that they were about as in *Protoceratops*.

the ceratopsian skull." In comparing *Protoceratops* with "the true Ceratopsia," Granger and Gregory (1923) say that in all the later forms there are supraorbital bones. Sternberg (1927, pl. 1) suggests a supraorbital in *Styracosaurus*, and figures (pl. 3) a fragmentary specimen of a young ?*Chasmosaurus* showing an element in front of, and suturally distinct from, the postorbital. This he calls the supraorbital. That the so-called supraorbital is present in the later Ceratopsia is not entirely without question. Hatcher, Marsh, and Lull, having studied the same specimens as did von Huene, make no mention of, nor do they figure the supraorbital in any of the specimens dealt with in their Monograph on the Ceratopsia. In *Brachyceratops*, the supraorbital is only supposed to have been present, and Sternberg has not recorded the possibility that the supraorbital in his immature ?*Chasmosaurus* might be the prefrontal, and what he supposes to be the fragment of the prefrontal might be a portion of the nasal. His interpretation seems most logical, however, since in *Chasmosaurus* the prefrontals unite in the median line to separate the nasals from contact with the frontals. It cannot be regarded, however, as the homologue of the true supraorbital. This element occurs only in some of the fishes and is unknown in any of the higher vertebrates.

In *Proceratops* there is a small bone over the front of the orbit which freely articulates with the prefrontal. To this Gregory and Mook (1925: 1) gave the name "palpebral bone" because they regarded it, and rightfully so, as the homologue of the "eye lid" bone of the Crocodilia, or of a form such as *Varanus*. They also intimated that these palpebral bones were homologous to the supposed "supraorbitals," for in designating the characters of the family Protoceratopsidae (p. 4), they state: "Freely articulating palpebral bones (supraorbitals) attached to the antero-superior corner of the orbits." They were followed in this by Sternberg (1927: 138) who says that in *Protoceratops* ". . . the supraorbital bones are freely articulating, palpebral bones . . ." Also, in describing the bones on the skull roof, Lull says (1933: 76): "I should interpret the pair of bones lying between the orbits on the dorsal surface of this skull as frontals, flanked in front and behind the orbit by the prefrontals and postfrontals respectively, the 'freely articulating palpebral bones' of Gregory and Mook representing the supraorbitals."

It seems probable that the "supraorbitals" in the later ceratopsian are the palpebral bones which have been incorporated in the structure of the superior orbital margin. This could be accomplished simply by the forward growth of the postorbitals and the mesial growth of the

prefrontals, which, of course, has happened. The postorbitals would then come in contact with the palpebrals and become suturally united with them. How such a transformation took place can be seen in homologizing the skull elements in *Protoceratops* and *Brachyceratops*. The unusual thecodont archaeosaur, *Sebecus icaeorhinus* Simpson (1937) from the Eocene of South America also suggests in a striking way, by the constriction across the frontals and the enlargement of the palpebrals, how this change could have been accomplished in the ceratopsians. (See PLATE 4.)

Postorbital

The postorbital is primitive in position and in form. Its position is for most part posterior to the orbit, where it forms the front of the narrow and still quite primitive postorbital-squamosal bar. It enlarges somewhat with age, arches quite pronouncedly, and becomes very rugose, thus foreshadowing the change that takes place in the later forms in which brow horn-cores are developed.

As in all ceratopsians, the ventral wing of the postorbital reaches the postero-inferior corner of the orbit. It is long and pointed in the young but with age becomes deeper and more bluntly united with the jugal below. The posterior wing in the beginning is short and fits into a rather deep notch in the squamosal. With age it extends backward and in the fully adult individuals it is considerably underlain by the ventral fork of the squamosal while the dorsal fork is abbreviated, although this feature is quite variable in some skulls. In skull Am. Mus. No. 6408, an immature female, the fork of the squamosal is almost lacking. The dorsal wing is short and well defined in the youngest specimens. In the older forms it is less defined and migrates forward over the posterior border of the orbit, and the postero-internal border of the whole bone becomes straightened.

There can be no question about the homology of the postorbital in *Protoceratops*, for, as in all of the ceratopsians, it receives on its under surface the vertical projection of the laterosphenoid.

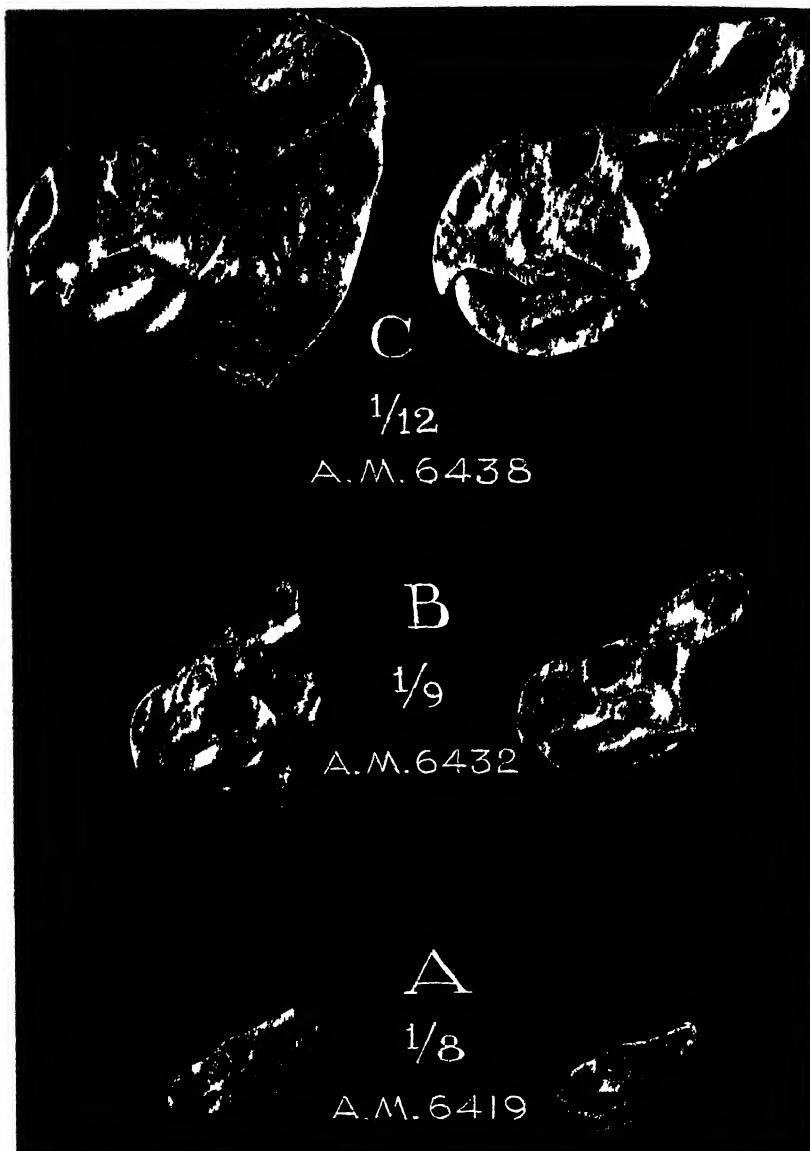
This element is of unusual interest, because, in its primitiveness and in the change it undergoes from youth to old age, it shows the first stage of what is perhaps the greatest transformation of any single element in the ceratopsian skull during the evolution of the group. From this primitive stage, the postorbital grows forward, unites with the enlarged palpebral, and thus eliminates the frontal and prefrontal from the margin of the orbit. At this stage, as shown by *Brachy-*



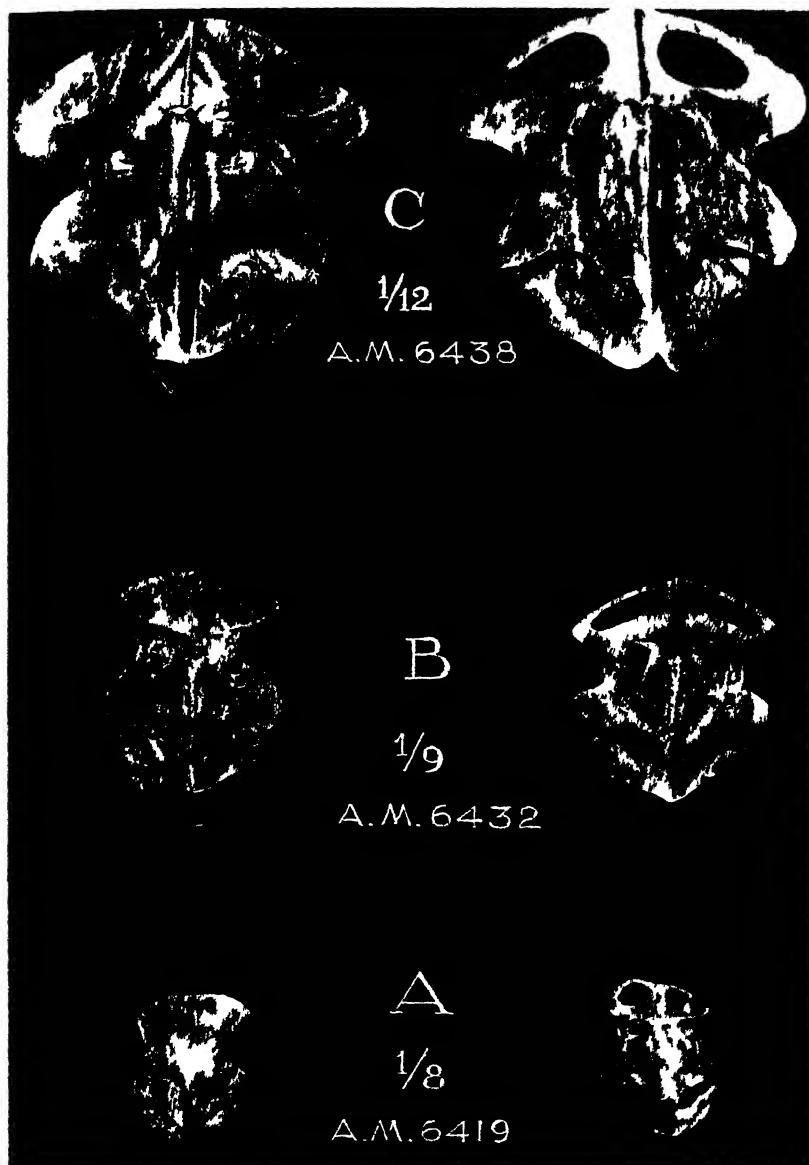
1/2

Sebecus trachynotus Simpson, a South American Eocene archosaur. Dorsal view of the skull, illustrating the development of the palpebrals and how they have invaded the skull roof.

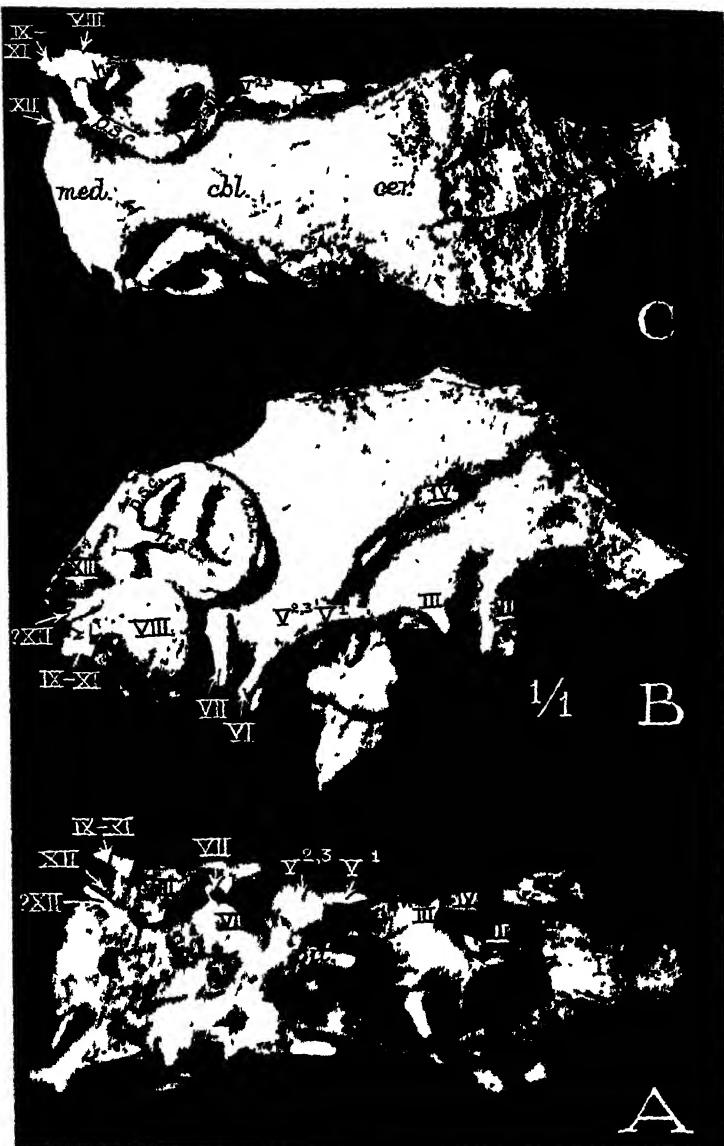
BROWN AND SCHLAIKER PROTOCERATOPS



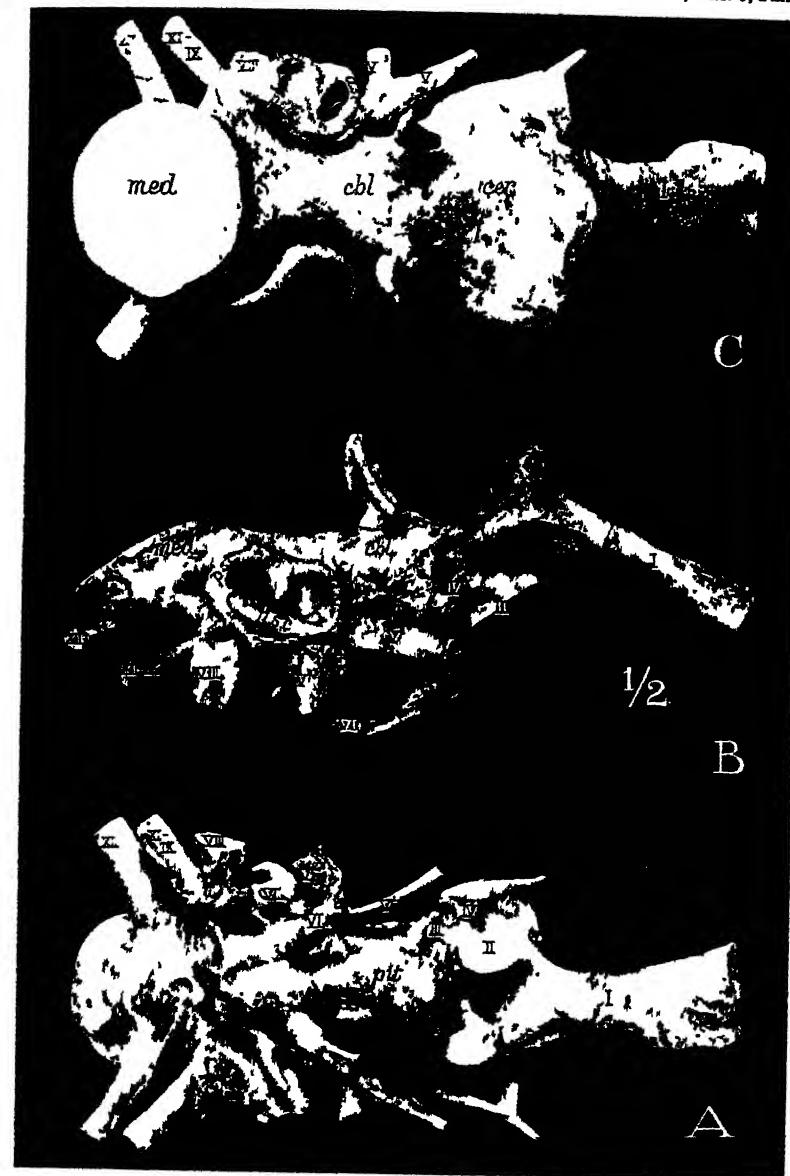
A series of skulls and jaws of *Protoceratops andrewsi* with restorations, lateral views
A, very immature individual B, young "male" C, old "male" Models by Sculptress
Georgia Mary Whitman



A series of skulls and jaws of *Protoceratops andrewsi* with restorations, anterior views
A very immature individual B, young "male" C, old "male" Models by Sculptress
Georgia Mary Whitman



Protoceratops andrewsi. Endocranial cast of a fully adult "male", Am. Mus. No. 6466, showing the main areas of the brain, the semi-circular canals, and the cranial nerves. A, ventral view. B, right lateral view. C, dorsal view. Cast by Otto Falkenbach.



Anchiceratops ornatus. Endocranial cast of the paratype, Am. Mus. No. 5259. A, ventral view. B, right lateral view. C, dorsal view. Cast by Otto Falkenbach.



Protoceratops andrewsi: Left lateral and dorsal views of the type skull and lower jaws,
Am Mus No 6251 One half natural size

ceratops, the dorsal surface protrudes to form an incipient brow horn-core. Apparently, the development of the brow horn-core is held in check, at least in the *Monoclonius-Triceratops* line, while the nasal horn-core proceeds to enlarge. This emphasis of the nasal horn-core development before that of the brow horn-cores is already established in *Protoceratops*. When a reduction of the nasal horn-core begins, the brow horn-cores then enlarge, and reach their maximum development

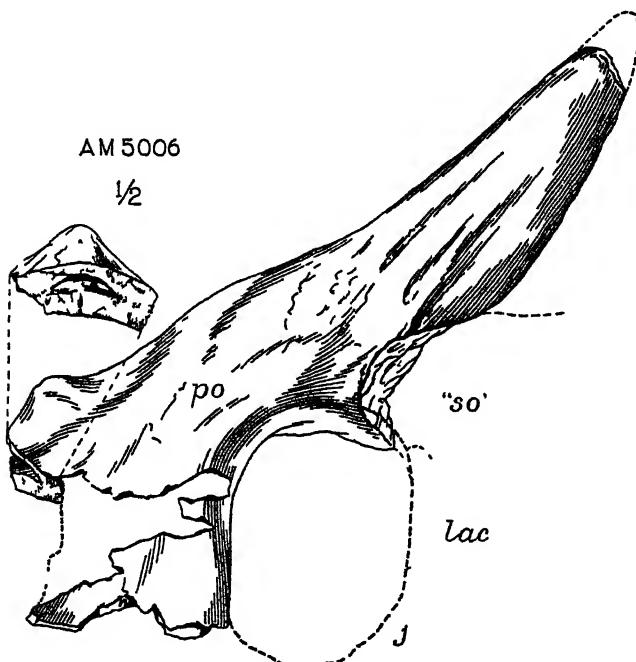


FIGURE 6. Postorbital bone of a very young *Triceratops* sp., from the Hell Creek beds of Montana showing the brow horn-core as an outgrowth of that bone. After Brown and Schlaikjer 1940a.

in *Triceratops* by the close of the Cretaceous. In *Triceratops eurycephalus*, the last of the known species, the brow horn-cores have reached extraordinary proportions and the nasal horn-core is reduced to a low base on which rests a small suturally distinct horn-core. (See FIGURES 5 and 6.)

The problem of origin and evolution of ceratopsian brow horn-cores has been fully treated in a previous paper by us (1940a).

Frontal

In the very young skull of *Protoceratops*, the frontals are primitive in form and in dimensions. They compose most of the skull roof between the orbits. Anteriorly they are elongated and form a wedge-shaped projection between the prefrontals and nasals that extends in front of the anterior margins of the proportionately much enlarged orbits. Posteriorly they deflect laterally to meet the lateral wings of the postorbitals, and the posterior contact with the parietals is broadly exposed on the skull roof. The dorsal surface is quite flat and is but slightly rugose.

With age, the following changes in the frontals take place. (1) The anterior projection becomes abbreviated and blunt, a change that results from the posterior extension of the nasals and prefrontals. (2) The exposure of the frontals on the superior borders of the orbits becomes restricted by the posterior growth of the prefrontals and the anterior growth of the postorbitals. (3) The posterior lateral projections are reduced, especially anteriorly, because of the anterior and mesial growth of the postorbitals. (4) The dorsal contact with the parietals becomes proportionately narrower, and the suture, which is straight at first, swings forward in the adult skulls. (5) There is a proportionate reduction in size.

In every one of these changes, the frontals foreshadow the morphology of these bones in all of the later Ceratopsia. In addition to these changes, there is one other which is of greater significance than any, and therefore deserves separate consideration. It concerns one of the most puzzling features of the ceratopsian skull,—the development of a secondary skull roof. Definite evidence of how and of what this secondary roof is formed has never been presented. In *Protoceratops* the very beginning of this secondary roofing is admirably shown. This new evidence, together with the evidence shown in *Brachyceratops*, as figured by Gilmore (1917, figure 3), in *Styracosaurus*, as figured and described by Sternberg (1927: 139–140, pl. 1), in *Monoclonius*, as seen in specimens in the American Museum (especially No. 5442), and in various species of *Triceratops*, gives the most probable answer as to the origin and composition of this unusual structure.

In the youngest of *Protoceratops* skulls, the frontals are nearly flat on the dorsal surface. Posteriorly at the median line they meet to form a small V-shaped projection wedged in between the front of the parietals. In those which we regard as probably male skulls, a slight depression soon appears in the young individual where the parietals and frontals unite on the dorsal surface. This depression is about

equally developed on the parietals and frontals and is bounded in front by a low edge that extends forward and inward from the postero-lateral margins of the frontals. At the median suture this edge extends posteriorly to form a short projection which is about as far forward from the frontal-parietal suture as the blunt anterior termination of the parietal crest is behind this suture. In the oldest skulls, as shown in Am. Mus. No. 6438 (FIGURES 7 and 8), this parieto-frontal depression becomes deeper, becomes more rounded in front, and becomes much more extensive—including nearly one-third of the dorsal surface of the frontals. In the male skulls, there is some variation. An occasional large skull may not have the depression as well developed as a smaller one. Specimen Am. Mus. No. 6425 is somewhat smaller than the largest (Am. Mus. No. 6438) yet the depression is scarcely formed. In none of the skulls, which we regard as females, is this depression developed, although the area is generally a bit concave and lacks the rugosity characteristic of the rest of the dorsal surface of the frontal.

It seems probable that this parieto-frontal depression of *Protoceratops* is the homologue of what is most frequently spoken of as the "postfrontal fontanelle," or what Sternberg (1927: 138) has more rightfully called the "frontal fontanelle," of the later ceratopsians. It shows, we think, the very beginning of the secondary roofing of the skull which is accomplished to such an extent in some species of *Triceratops* that the fontanelle opening is completely obliterated.

The transitional stages between the open parieto-frontal depression of *Protoceratops* and the nearly, or completely, closed over secondary skull roof of *Triceratops* are admirably shown by *Brachyceratops* and *Monoclonius*. In *Brachyceratops* the depression extends even farther forward than in *Protoceratops*. It has become somewhat deeper, and with the enlargement of the postorbitals and the appearance of postorbital horn-cores, it is becoming rather constricted laterally. The only recorded specimen of *Brachyceratops* is an immature individual and it is entirely possible that all of these characters may be even more emphasized in the adult stage. Such is the case with *Protoceratops* and it seems reasonable to expect the same in *Brachyceratops*.* Also, mainly as a result of the changes in the postorbitals, the frontals have become eliminated from the orbital borders and the frontal-postorbital sutures are in a more antero-posterior position. Anteriorly they extend along the inner flanks of the post-

* Recently Gilmore (1939: 12) has described the frill of an adult. The characters displayed in the portion which unites with the frontals substantiates this suggestion.

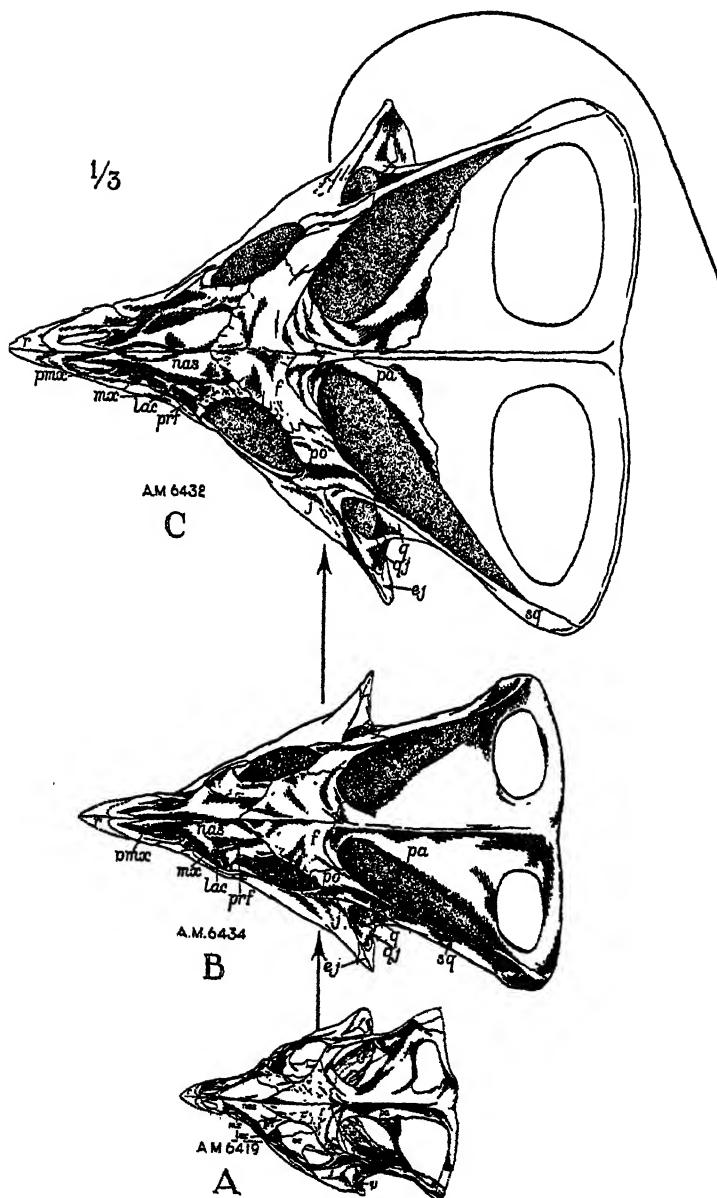
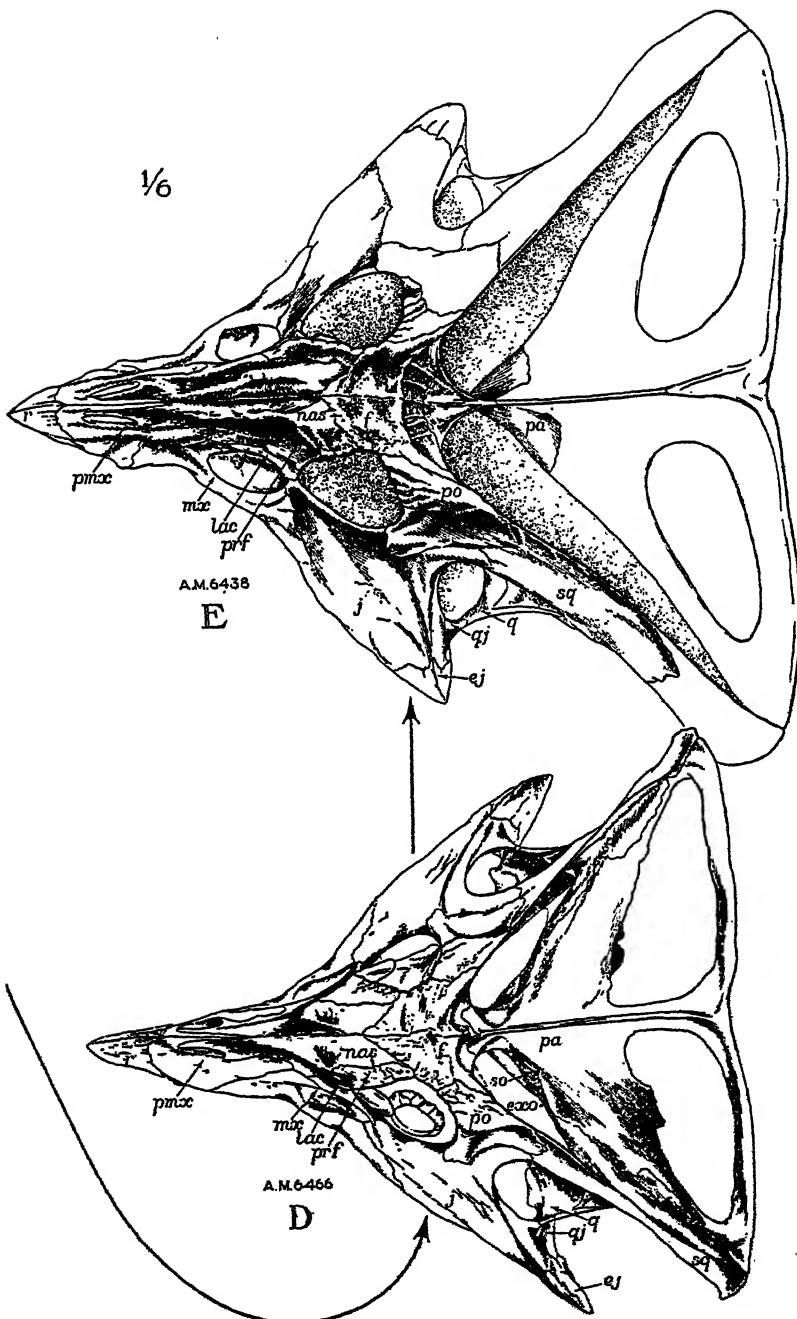


FIGURE 7. Series of *Protoceratops* "male" skulls showing the development from a very immature (A) to an old (C) individual.



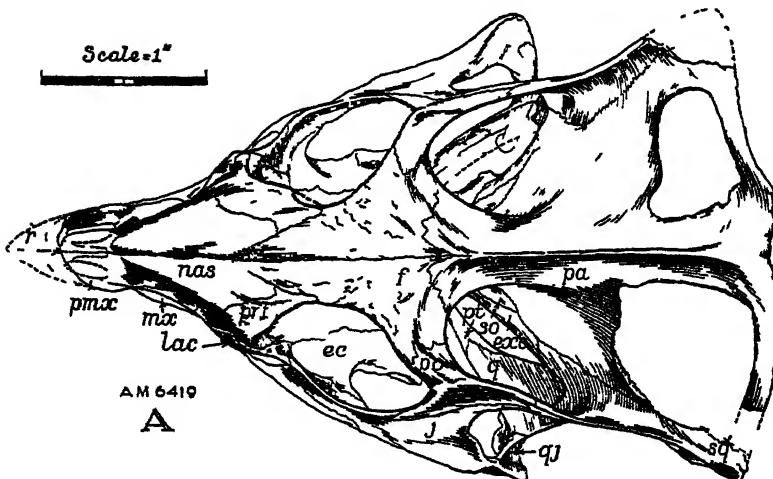
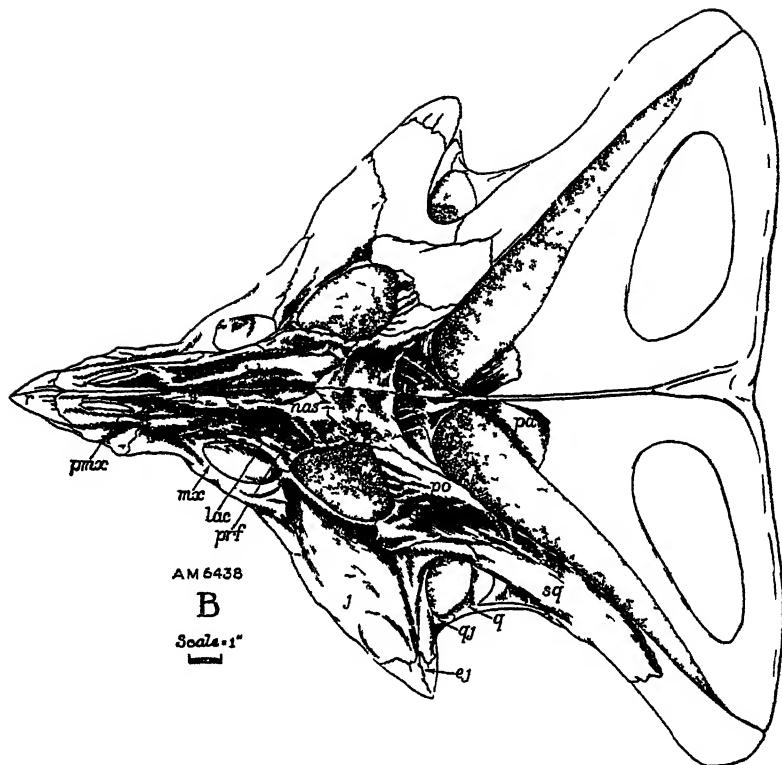


FIGURE 8 Two *Protoceratops* skulls, dorsal views, displaying the marked growth change
A, very immature individual B, old individual Not to scale.

orbital horn-cores and then meet medially to form the usual anterior projection. At this stage the prefrontals have come in contact with the postorbitals and have been eliminated from the orbital borders by the palpebral bones, or "supraorbitals," which have been taken over into the skull roof and suturally united with the prefrontals and postorbitals. The blunt anterior termination of the parietal crest in *Brachyceratops* has become broadened and heavy, thus indicating the beginning of the concentration of the capiti-mandibularis insertion in this region of the crest.

In *Monoclonius*, as in *Styracosaurus*, the parieto-frontal depression has become considerably deeper and so laterally compressed that its sides are now parallel and are somewhat overhanging. Along with this change, the antero-lateral margins of the parietals have migrated inward, and upward to a level almost the same as that of the lateral portions of the frontals, which have remained on the surface of the skull roof. As in *Brachyceratops*, the anterior projection of the frontals also remains on the surface, and the relationship of the frontals to the other roof bones is about the same as in that genus.

Chasmosaurus, in which the brow horn-cores are considerably developed, shows a stage in this secondary roofing of the skull that is structurally intermediate between that of *Monoclonius* and *Triceratops*. Between the horn-cores, the fontanelle is quite constricted, and while the troughs leading from the fontanelle across the antero-lateral margins of the parietals to the temporal fossae are shallow, they are still well emphasized. In *Triceratops*, there is a considerable amount of variation in the fontanelle region. In the skull (No. 5116) of the mounted composite skeleton in the American Museum, which seems nearest to *T. elatus*, the opening is fairly large and the troughs leading to the temporal fossae are quite well developed. In the skull of *T. hatcheri* (U.S.N.M. 2412) the troughs are present but do not reach the temporal fossae. In skull No. 907, referred to *T. serratus*, of the American Museum collection, the opening is still present, but the troughs are almost completely obliterated. And in the type skull of *T. prorsus* (Y.P.M. 1822) the secondary roof of the skull is complete with no vestige of the opening whatsoever.

On the basis of our present knowledge of the ceratopsian skulls, it seems therefore, that the secondary skull roof has resulted from the deepening, the elongation, and the gradual closing over of the parieto-frontal depression that is already well established in the adult *Protoceratops*. As suggested in the development of *Protoceratops*, this change certainly seems correlated with a need for strengthening as the

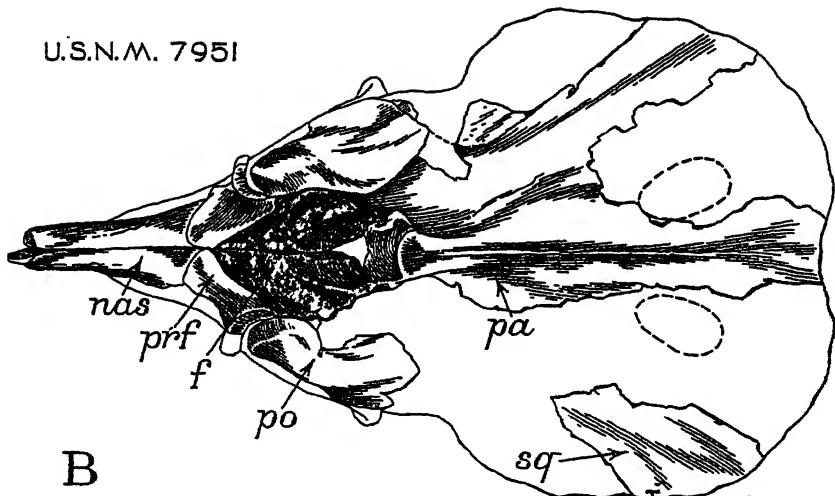
frill widened and became more erect and the forward stresses from the jaw muscles that were transmitted through the crest and along the squamosals became focused in the frontal area (see FIGURE 10), and with the appearance and growth of the brow horn-cores (see FIGURE 5).

During this change, the frontals, of course, have undergone remarkable modification, and a considerable portion of them has been eliminated from the dorsal surface of the skull. Just how much has been excluded from the skull roof is a question which has given rise to much discussion. Von Huene (1911: 157) shows the frontals of *Triceratops elatus* entirely excluded from the skull roof and lying under what he called the lachrymals (really prefrontals). The evidence at hand certainly militates against such a conclusion. In all of the earlier ceratopsians,—*Protoceratops*, *Brachyceratops*, *Monoclonius* (*Centrosaurus*), *Styracosaurus*, and *Chasmosaurus*—in which the boundaries of the frontals are known to be suturally distinct, the frontals are on the surface anteriorly and along the margins of the fontanelle. In *Triceratops*, there is some question, however, as to how much of the frontals, if any, forms the secondary roof. There is a possibility that, with the enormous enlargement of the brow horns, the posterior portions along the margins of the fontanelle were crowded inward and completely folded under. This also seems improbable for the following reasons. First, in the earlier forms such as *Monoclonius* (*Centrosaurus*) the frontals are heavy, rugose, and well developed along the margins of the fontanelle. Second, in no *Triceratops* skull are the frontals shown to be completely buried beneath the secondary roof. Third, along the entire median surface of the postorbital of the immature *Triceratops* (Am. Mus. 5006), shown in FIGURE 6, there is a deep frontal suture, which proves that in *Triceratops* the frontals are thick heavy elements, exposed on the surface, along their lateral margins where they are in contact with the postorbitals.

This evidence seems to show that in the highly specialized *Triceratops*, in some species of which the fontanelle is entirely obliterated, the frontals occupy virtually the same position and have essentially the same relationship to the other cranial elements as in *Monoclonius*.

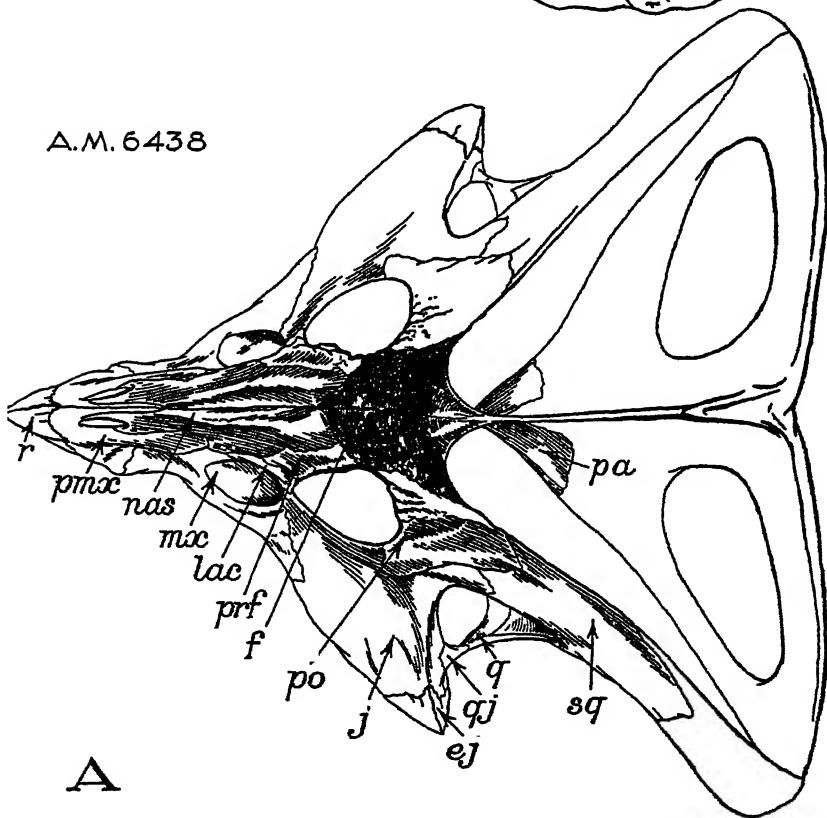
In his detailed description of *Brachyceratops*, Gilmore (1917: 10, 11) described the paired elements between the postorbitals on the skull roof as the postfrontals and, following von Huene's conclusion that the frontals in *Triceratops* are entirely excluded from the dorsal surface, considered the frontals to be buried beneath them. In the light of the vast amount of material that has been discovered since Gilmore's description, this now seems to be entirely improbable. In the first

U.S.N.M. 7951



B

A.M. 6438



A

FIGURE 9. Dorsal views of two ceratopsian skulls showing the early stages in the development of the parieto-frontal depression. A, *Protoceratops andrewsi*, an old "male." B, *Brachyceratops montanensis*, modified from Gilmore. Not to scale.

place, there is no evidence whatsoever that there were any elements under the so-called "postfrontals" of *Brachyceratops*. Secondly, the formation of a secondary skull-roof is a structure peculiar to the later and more highly specialized ceratopsians, and is completely established only among the most highly specialized species of those forms. It seems improbable, therefore, that a form so almost completely primitive as *Brachyceratops* should be so highly specialized in this one character, particularly when the closely associated characters are so archaic. Thirdly, the elements which Gilmore calls the "postfrontals" are in the same position, hold exactly the same relationship to the other cranial elements, and bear the identical postero-median, or parieto-frontal, depression as in *Protoceratops*; and, as discussed above and as is shown in FIGURE 9, they represent an ideal intermediate stage in the development of the secondary skull roof between that genus and *Monoclonius*.

Parietal

Perhaps no element in the ceratopsian skull has been so variously designated and has received so much comment as the middle portion of the frill. Hatcher, Marsh, and Lull in their monograph on the Ceratopsia (1907: 19) interpreted it as the fused parietals. Hay (1909: 97) regarded it as composed of the supratemporals or possibly nuchal bones homologous to those found in the Crocodilia. Von Huene (1911) considered the anterior portion immediately behind the frontal fontanelle as the fused parietals and the remainder as the fused dermosupraoccipitals. Gilmore, in his description of *Brachyceratops* (1917: 11), referred to this element as the dermosupraoccipital or interparietal—a view he continued to hold, though hesitatingly so, in a later paper (1930: 36). This confusion has resulted mainly from a misidentification of the cranial elements immediately in front of the frill, and a mistaken identification of cracks as sutures in adult skulls of *Triceratops*.

The magnificent collection of *Protoceratops* skulls showing the gradual development from a very young to a very old stage displays most convincingly that the central element of the frill is unquestionably composed of the fused parietals (see FIGURE 7). Even in the youngest skulls they are almost completely fused, showing only a slight indication of a suture for a very short distance just behind where they are separated by a blunt projection of the frontals. This early fusing of the parietals is a necessity since they function as an anchorage for the great capiti-mandibularis muscle masses.

On the dorsal surface they are in extensive contact with the frontals.

This contact continues laterally for a short distance where the parietals come in contact with the laterosphenoids, the supraoccipitals, the exoccipitals, and finally the squamosals, with which they are in contact, throughout the lateral portions of the frill. Postero-inferiorly they unite with the supraoccipitals which become pinched out from between them and the exoccipitals about one-third of the distance out from the medial line.

It is obvious, therefore, that the median frill element not only has the position expected of the parietals, but has the correct relationship to the other cranial elements. Furthermore, it forms the postero-superior portion of the brain case. In the young skulls the suture between it and the frontals is about midway over the brain and in the older skulls shifts posteriorly somewhat. In the later ceratopsians this anterior portion of the parietals, that forms part of the brain-case, becomes buried by the secondary roofing-over of the skull in a manner which has been fully described above.

In commenting on the question of the parietals forming the middle part of the frill of *Protoceratops*, Gilmore (1930: 36) says, ". . . although it appears to represent that bone in *Protoceratops*, it certainly cannot be the parietal in American horned dinosaurs, as evidenced by the juvenile *Brachyceratops* and other ceratopsian skulls in the National collections." The main difficulty with Gilmore's interpretation of the frill of *Brachyceratops* lies in his mis-identification of the frontal bones, which is discussed above. The fact is that in *Protoceratops* the parietals have identically the same relationships with the other cranial bones as does the middle frill element of the later ceratopsians.

The very immature skull of *Protoceratops* shows an early stage in the development of the frill. The postero-superior portion of the skull has just begun to be drawn out posteriorly and acts as a scaffolding for the large capiti-mandibularis masses. At this stage, the frill is, therefore, short, and only slightly expanded posteriorly, and the temporal openings are proportionately large and rounded in form. The crest is but a low ridge, and the median surface of the frill is more or less on the same plane as the dorsal surface of the frontals. The fenestrae are quite large and rounded, and are bounded posteriorly by the thin and narrow marginal portions of the parietals. With age, the frill elongates and becomes greatly expanded. The temporal openings are more laterally constricted but are more drawn out. The crest of the frill becomes very heavy and high, and the frill becomes steeply inclined to the plane of the frontals. The parietal fenestrae are somewhat variable in size and form in the older skulls, but in the fully

adult they are definitely reduced and on their posterior margins the parietals are broadened and much heavier.

As suggested earlier, the proportionate widening of the frill results in the modification of the parieto-frontal area. In the young skull, the sides of the frill, along the outer margins of the squamosals, are

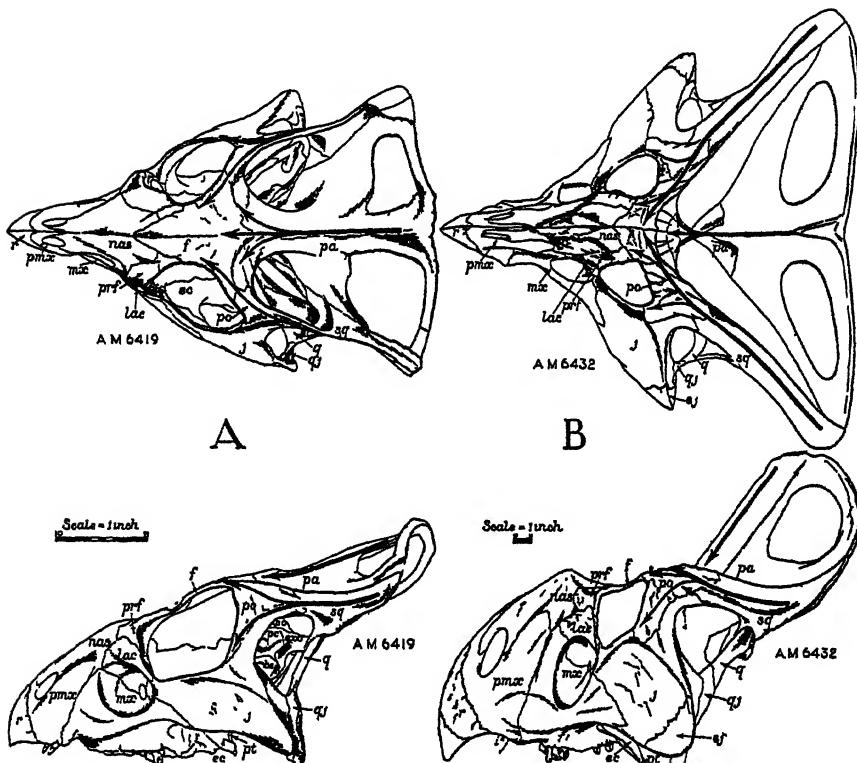


FIGURE 10. Lateral and dorsal views of a very immature skull (A) and old skull (B) of *Protoceratops andrewsi* showing how the parieto-frontal depression develops when the stresses from the capiti-mandibularis muscle masses become concentrated in that area. Heavy and light arrows mark amount and direction of stress.

almost parallel, being only slightly deflected outward posteriorly. In this stage, the stresses from the pull of the capiti-mandibularis masses are transmitted forward and downward. The forward stresses are transmitted through the median portions of the parietals to the frontals, and laterally through the squamosals to the postorbitals where they are distributed around the orbits, although for the most part ventrally. As the frill widens, the squamosals are widely dis-

placed posteriorly and the postero-internal borders of the postorbitals are straightened. The postorbital-squamosal bars, therefore, become directed inwardly towards the posterior area of the frontals. This results in a concentration of the lateral forward stresses in this area—especially since the upward and forward growth of the postorbitals and the upwardly inclined squamosals cause most of the lateral stresses to be directed up over the orbit instead of below it as in the very young skulls. In addition, since the frill has become steeply inclined, the median forward stresses are directed more downward than forward and likewise become concentrated at the posterior area of the frontals.

When this concentration of stresses took place, there seems to have been a need for strengthening in the parieto-frontal area of the skull roof. The formation of the parieto-frontal depression (see above), with its reinforced margins, supplied the need. In the female skulls, however, this depression is not nearly as well developed, but these skulls remain quite primitive in this area. The angle of the postero-internal borders of the postorbitals is not completely lost even in the adult skulls. Thus the lateral forward stresses are directed more anteriorly along the superior margins of the orbits. Also, the frill never becomes steeply inclined to the plane of the frontals. A flattening and a somewhat down-warping of the parieto-frontal area seems to have supplied sufficient strengthening. (See FIGURES 11C and 12B.)

The parieto-frontal depression, or fontanelle, once established, its further development and closing over in the later ceratopsians were accomplished by the enlargement of the postorbitals which proceeded with the appearance and development of the brow horn-cores. With this transformation, came the necessity for a more solid construction in the frill region. In the later ceratopsians, therefore, the capitomandibularis masses become concentrated anteriorly and lose their great lateral extent posteriorly. As this takes place, the high frill crest, so characteristic in the adult *Protoceratops*, is lost, and the whole frill becomes far less steeply inclined to the frontal area.

Squamosal

The squamosal of *Protoceratops* is very primitive in its position and in its general form. It is situated high on the side of the skull and its descending wing does not come into contact with the jugal and (or) quadratojugal—a feature characteristic of all the known later ceratopsians in which these elements are preserved. Together with the postorbital and the slight ascending wing of the jugal, it forms a narrow and distinct postorbital-squamosal bar above the large lateral

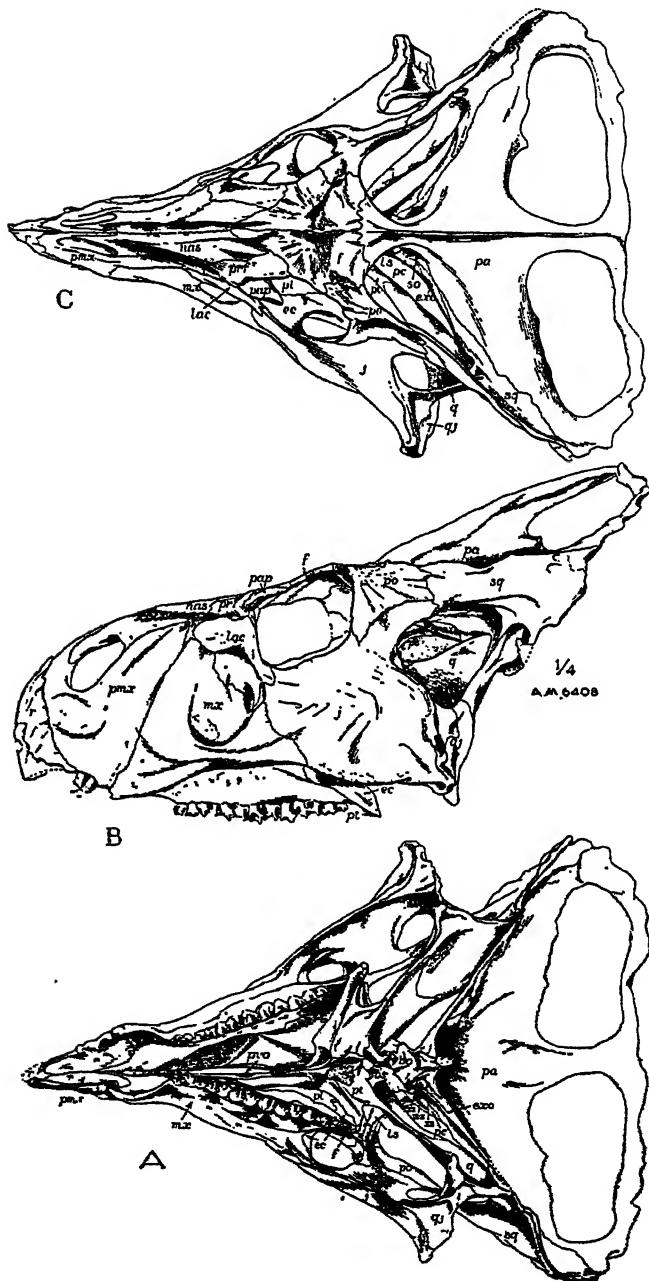
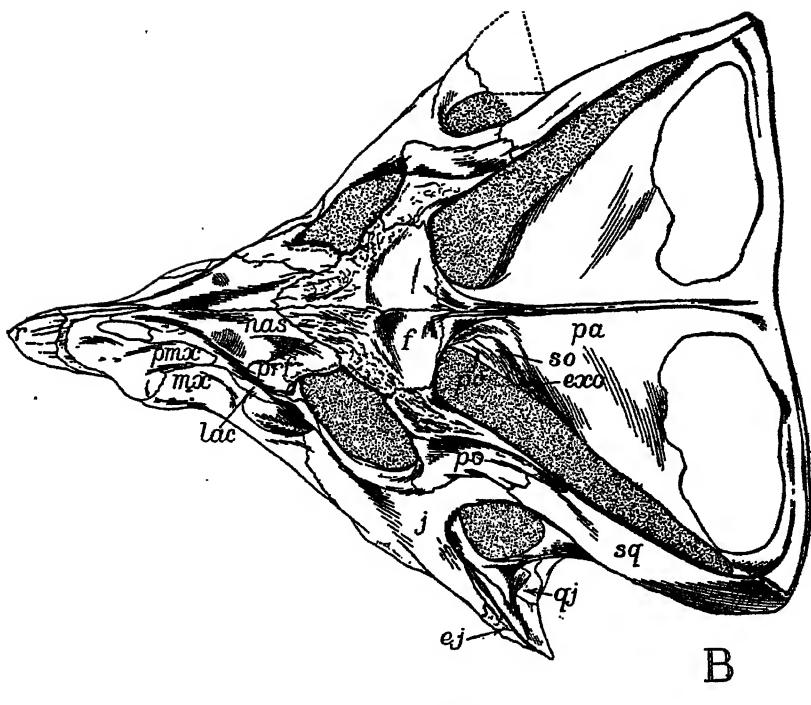


FIGURE 11. *Protoceratops andrewsi*. Ventral, lateral, and dorsal views of a fairly large, young adult "female" skull.



A.M. 6429 1/4

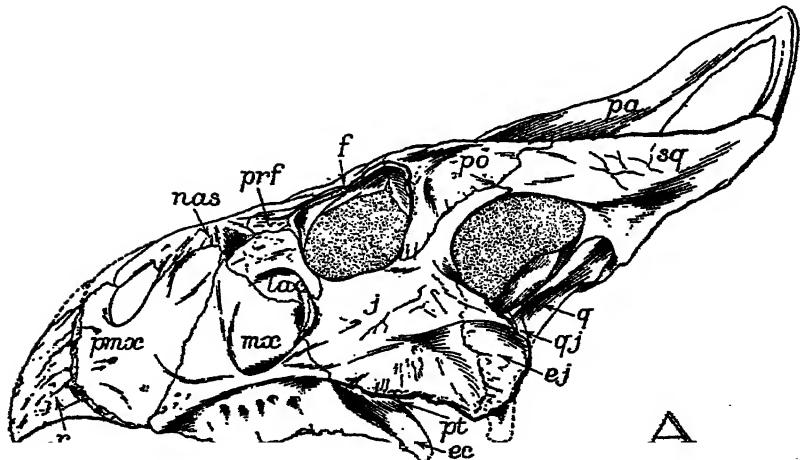


FIGURE 12. *Protoceratops andrewsi*. Lateral and dorsal views of an adult "female" skull.

temporal opening. It forms the superior postero-lateral part of the skull and is braced between the quadrate and the parietal, and between the postorbital and the ascending wing of the jugal and the parietal.

It is composed of three main branches, an antero-posterior, a ventral, and a lateral. In the young and in some of the fairly adult skulls, the antero-posterior branch is quite horizontal in position. In the old skulls, however, in which the postorbital is proportionately enlarged and arched upward, this branch rotates so that it becomes directed upward anteriorly. This inclination of the antero-posterior branch of the squamosal is further exaggerated in the later ceratopsians. In these, the dorsal margin of this branch, at least the anterior part of it, is inclined with about the same steepness as the posterior profile of the brow horn-core.

The antero-posterior part of the squamosal is proportionately narrow and short in the beginning. It deepens with age and as the frill enlarges and grows posteriorly, it is dragged back and therefore becomes elongated. Naturally this elongation is reflected in the portion posterior to the ventral wing, but anterior to this wing the part of the squamosal included in the postorbital-squamosal bar also becomes proportionately elongated. These changes also affect the form and the position of the lateral temporal opening. In the young skull this opening is nicely rounded in front, and the straight posterior margin is formed mainly by the nearly erect quadrate. With age, the dorsal border becomes flattened, the posterior margin becomes steeply inclined forward, the postero-superior corner becomes pointed, and the entire fossa shifts slightly backward with respect to the orbit. Although the lateral temporal opening is rather variable in its proportionate size, there is a tendency for it to be reduced with age.

The anterior margin of the squamosal that is in contact with the postorbital and the jugal also becomes modified with age. At this contact, the squamosal presents two forwardly directed projections—one above and one below the posterior point of the postorbital. In the young skull, the superior projection is more pronounced and overlaps the postorbital. With age, the ventral projection increases in size and extends under the postorbital for a considerable distance, while the dorsal projection is comparatively short and blunt. This probably is due mainly to the enlargement of the postorbital, especially its growth posteriorly.

There is not the slightest bit of evidence on any skull that epoccipitals were present on the margin of either the parietal or the squamosal.

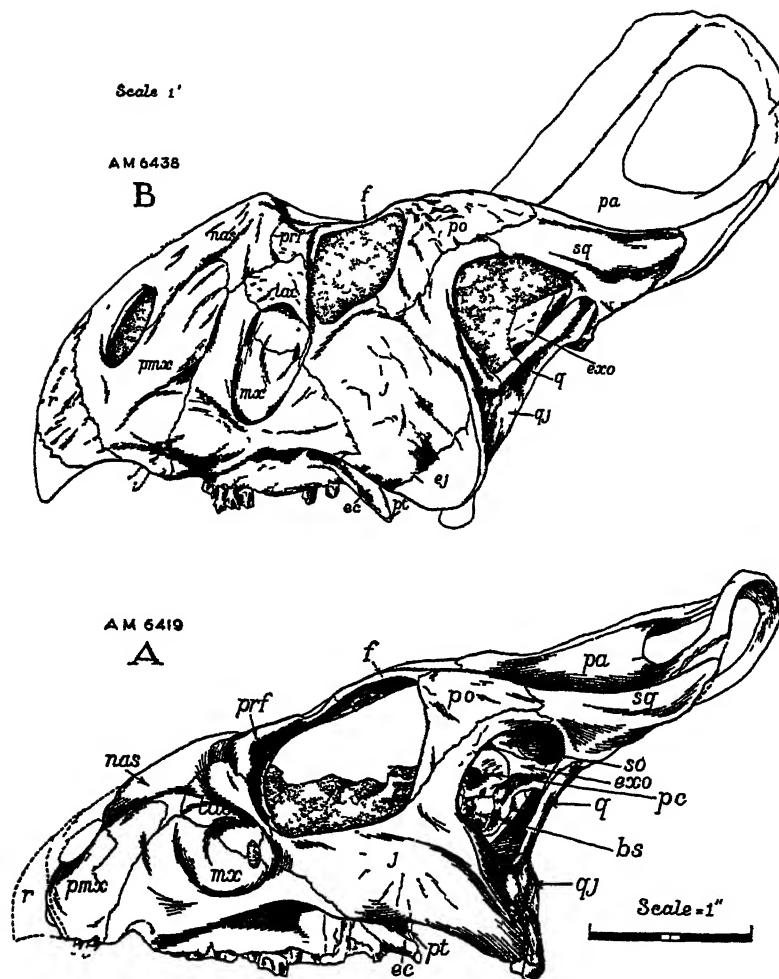


FIGURE 13. Two *Protoceratops andrewsi* skulls, lateral views, displaying the marked growth changes. A, very immature individual. B, old "male." Not to scale.

On the postero-external surface of the ventral branch of the squamosal there is a marginal pocket which receives the external end of the exoccipital. On the front of this branch another pocket is formed into which the superior branch of the quadrate is buried firmly. In addition, a blunt spur of the squamosal extends down for a short distance in front of the quadrate just in from the external margin. With age, the distance between it and the quadratojugal lessens, although it never reaches that bone as it does in the later ceratopsians in which the

lateral temporal opening becomes proportionately smaller, and the anterior area of the squamosal rotates so that it is directed steeply upward. Thus in *Protoceratops* the quadrate has already become deeply buried in the ventral portion of the squamosal in a characteristic manner that is peculiar to all ceratopsians. By this arrangement, most of the vertical stresses from the quadrate-articular joint are disseminated around the outer margin of the frill, and counteracting these are the downward stresses from the pull of the jaw muscles.

From the base of the ventral branch of the squamosal, the lateral branch extends inwardly on the front of the exoccipital, but never quite reaches the supraoccipital. Upward and backward from this point, the squamosal forms a wide lateral margin of the frill along which it overlaps considerably the outer border of the parietal. This relationship with the exoccipital and parietal is retained throughout the evolution of the Ceratopsia.

The more outstanding changes that take place in the squamosal from the young to the adult *Protoceratops* are as follows:

1. The antero-posterior branch rotates from quite a horizontal position so that it becomes directed upward anteriorly.
2. This branch elongates and increases in depth.
3. The ventral projection becomes more closely approximated to the quadratojugal.
4. The lateral temporal opening, of which the posterior part of the upper border is formed by the squamosal, has a tendency to become proportionately smaller.

These changes are carried out even more strikingly in the later ceratopsians in which the squamosal becomes a highly modified element. This is shown in the following salient features.

1. The anterior portion becomes directed steeply upward, especially in those forms with the much enlarged brow horn-cores.
2. It becomes very deepened and plow-share-like in general form, and increases in length, particularly in the *Chasmosaurus-Torosaurus* line.
3. The ventral projection assumes contact with the quadratojugal and the contact with the jugal above the lateral temporal opening becomes very extensive.
4. With the enlargement of the squamosal-jugal contact, the lateral temporal opening shifts downward far below the level of the inferior border of the orbit, and is much reduced.
5. With the reduction of the supratemporal opening, the posterior portion of the lateral branch of the squamosal merges with the posterior

part of the antero-posterior branch, and the contact with the parietal becomes extensively exposed on the lateral surface of the frill.

6. Processes or epoccipitals are nearly always present on the lateral margin of the squamosal.

Jugal

The jugal is proportionately larger than in any of the other known ceratopsians. It has essentially the same relationship with the other elements on the side of the skull except that on its antero-internal side it is extensively in contact with the ectopterygoid (transverse bone). In the later forms, at least in *Styracosaurus* and *Triceratops*, the ectopterygoid is much reduced and not in contact with the jugal. Changes in proportions from the young to the adult stage are rather pronounced. The principal changes are as follows (see FIGURES 4, 8, and 13):

1. Primarily as a result of the deepening of the face, the contact with the lachrymal becomes more extensive, and the contact with the maxillary, which in the beginning is quite oblique, becomes more erect.
 2. With the proportionate reduction of the size of the orbit, the orbital border formed by the jugal is restricted and is confined more to the antero-inferior border of the orbit.
 3. The depth of the jugal under the orbit is much increased.
 4. With age, the jugal rotates from a quite horizontal to a rather vertical position and becomes proportionately broader.
 5. The long and narrow ascending wing of the young skull becomes proportionately shorter and heavier.
 6. When seen from above, the jugals set in a rather vertical plane in the young skull, but with age, they flare out below so that in the adult skull nearly all of lateral surfaces come into view.
- Nearly all of these changes in the jugal are those which are featured in the evolution of the skull in the later ceratopsians.

Epijugal

The epijugal is somewhat transversely compressed, although it is quite conical in its general form. It is proportionately large, rather pointed, and fits over the distal end of the jugal, overlapping a part of the quadratojugal suture. The surface is very rugose with the grooves and ridges directed towards the apex. This suggests that in life this element was covered with a horny sheath.

Quadratojugal

The quadratojugal is not in contact with the squamosal. This feature is distinctive of *Protoceratops*, for in all other known ceratopsians, the inferior border of the lateral temporal opening is formed by a union of the squamosal with the quadratojugal, as in *Triceratops*, or with the jugal, as in *Monoclonius*. Postero-ventrally the quadratojugal extends more around on the quadrate than in the later forms. Also, posteriorly the quadratojugal-jugal suture is straight in the later forms, whereas, in *Protoceratops* it extends upward, curves inward abruptly and then continues upward again. The interno-ventral extension is quite a distance above the articular surface of the quadrate. In the more advanced ceratopsians, as in *Triceratops*, it extends down to the very margin of the condylar surface of the quadrate, although never forming part of the jaw-articulation. There is an antero-inferior projection that extends forward for quite some distance along the inside of the ventral margin of the jugal. This character seems unique for *Protoceratops* since in the later types, the anterior border of the quadratojugal curves slightly forward and upward from the ventral margin.

Quadrate

The lateral wing of the quadrate is well marked off from the main shaft and is very much extended. It extends inward to a point considerably mesial to the line of the tooth-row and gives the quadrate a width that is over two-thirds its depth. This lateral wing becomes abruptly pointed at its extremity and overlaps extensively on the front of the postero-lateral wing of the pterygoid. On its infero-posterior surface the lateral wing of the pterygoid branches distally. The upper of these extends along near the upper border of the quadrate almost to where the latter is nearly in contact with the exoccipital. The lower branch is shorter and fits against the quadrate in what seems to be the beginning of a very shallow pocket.

Dorsally, the lateral wing merges with the main shaft to form a rather pointed projection that fits deeply into a pocket in the squamosal. It is wedged firmly in this pocket by a ventral projection of the squamosal that extends downward just inside of the main shaft on the supero-anterior surface, and by another ventral projection that extends downward on the infero-posterior surface. The latter intervenes between the quadrate and the exoccipital so that these two bones do not come in contact with one another. This character is unique in *Protoceratops*. In all other known forms the exoccipital is firmly united with the quadrate and the intervening posterior ventral projec-

tion of the squamosal is incipient. In the earlier forms, as in *Monoclonius*, the union of these elements is not as extensive as in the later types such as in *Triceratops*.*

The external surface of the quadrate is straight and not bowed outward as in *Triceratops*. Although the articular surface, and the extent of the externo-anterior projection on the front of the quadratojugal are about the same as in that genus. Since the quadratojugal does not extend down to the margin of the articular surface, there is a neck-like area developed just above that surface. In relation to the horizontal plane of the skull, the position of the quadrate is rather erect.

In a number of features, the quadrate of *Protoceratops* is quite different from that of the later ceratopsians. The more important of these are the following:

1. The lateral wing is more clearly demarcated from the main shaft, it is much more extended mesially, and on its infero-posterior surface there is no well formed pocket or notch for reception of the overlapping process of the pterygoid.

2. The quadrate is not in contact with the exoccipital. This contact is prevented by the posterior ventral projection of the squamosal which extends downward on the infero-posterior surface of the quadrate and intervenes between the two bones. In all other known ceratopsians the posterior ventral projection of the squamosal is abbreviated and the quadrate and exoccipital are extensively in contact with one another.

3. The external surface is straight, and the quadratojugal does not extend down to the articular surface.

4. In relation to the horizontal plane of the skull, the quadrate has a much more erect position than in any of the known later ceratopsians.

In all of these characteristics, the quadrate of *Protoceratops* is much more primitive.

Exoccipital

The exoccipitals are relatively long and slender laterally. In *Protoceratops* they have not as yet become broadened and abbreviated by the enlargement of the squamosals and parietals as in the later ceratopsians in which they have become large, heavy buttresses in the ventral architecture of the much enlarged frill. They unite ventrally

* In their figure of the quadrate of *Triceratops flabellatus*, Hatcher, Marsh, and Lull (1907: 23, fig. 17 A) have failed to show the sutural surface for contact with the exoccipital, which is located just above the surface overlapped by the pterygoid, and which extends across to the outer margin of the quadrate. The relationship of these elements is well shown on plate 33 of their Monograph.

to form the lower margin of the large, round foramen magnum, and thus eliminate the basioccipitals from entering into the formation of the margin of that opening. This is a characteristic of all known members of the suborder. They are unique, however, in not uniting above the foramen. The supraoccipital intervenes.

The lateral extension is but slightly expanded distally and is shallowly embedded in the squamosal only on the upper margin. The rest of the external margin is rounded and extends out to the lateral border of the skull behind the upper portion of the quadrate from which bone it is separated, as was explained earlier, by the posterior ventral projection of the squamosal. The superior margin of the lateral extension is firmly in contact with the parietal for only a very short distance just beyond the distal end of the supraoccipital. Beyond this, the median extension of the squamosal intervenes between the two bones so that the parietal has a slight contact with the rest of the lateral extension on the posterior surface only.

In the very immature skull the exoccipitals do not enter into the composition of the condyle, although there is a slight articular surface along their margin where they come together and jut out slightly above the condyle. In the adult, this margin merges with the basioccipital and forms the superior and supero-lateral portions of the condyle. Together, however, the exoccipitals form only about one-third of the condyle, whereas in the later forms, such as *Torosaurus*, according to Hatcher (Hatcher, Marsh, & Lull 1907: 17), the basioccipital and the two exoccipitals contribute about equally to its composition.

On the postero-ventral surface at the base of the lateral extension there are three foramina. The upper and most posterior of these enters the brain cavity just inside the foramen magnum. This probably was the exit for the twelfth nerve. Another foramen, about of the same size, is situated just below and slightly external to that for the twelfth nerve. It passes forward through the ventral portion of the exoccipital and enters the fenestra ovalis. This undoubtedly is the foramen lacerum posterius and probably transmitted the ninth and tenth nerves. The third foramen is somewhat below and about on a line between the other two, although it is rather closer to the foramen lacerum posterius. So far as is known, *Protoceratops* is the only ceratopsian to have this foramen. In all other genera the exoccipital has but two foramina. Its entrance directly into the brain cavity in front of the exit for the hypoglossal nerve, and its proximity to the foramen lacerum posterius, suggest that it gave passage to the eleventh

nerve. The absence then of this foramen in the later known ceratopians simply means that it has become confluent with the foramen lacerum posterius. Another, and perhaps more plausible, explanation, however, is that it was the exit for a separate branch of the hypoglossal nerve. In this case, in the later ceratopians it has become eliminated.

The relationship of the exoccipital to the proötic is about as in the other forms except that the postero-lateral wing of the proötic extends farther out on the antero-lateral surface of the exoccipital in *Protoceratops*.

A comparative list of the more important changes in the ceratopsian exoccipitals from *Protoceratops*, a primitive form, to *Triceratops*, an end-member of the group, are as follows (also see FIGURE 14):

Protoceratops

1. They do not unite above the foramen magnum to exclude the supraoccipital from the border of that foramen.
2. Unite below the foramen magnum and in the adult, form approximately one-third of the occipital condyle.
3. Lateral extensions relatively long and slender, and extend out to the lateral margin of the skull.
4. Distal ends of lateral extensions only slightly expanded, and shallowly embedded in the squamosals only on their upper margins.
5. Not in contact with the quadrate.
6. Firmly in contact with the parietals only for a short distance at distal ends of the laterally extensive supraoccipital.
7. Foramen for the eleventh or second branch of the twelfth, nerve distinct.
8. Sutural surfaces for the proötics extensive.

Triceratops

1. They unite above the foramen magnum and exclude the supraoccipital from the border of that foramen.
2. Unite below the foramen magnum and form approximately two-thirds of the occipital condyle.
3. Lateral extensions short, broad, heavy buttresses in the ventral architecture of the frill, and do not extend out to the lateral margins of the skull.
4. Distal ends of lateral extensions very much expanded, and deeply embedded in the squamosals throughout.
5. Firmly and extensively in contact with the quadrate.
6. Firmly in contact with the parietals throughout the superior margins beyond the much abbreviated lateral margins of the supraoccipital.
7. Foramen for the eleventh, or second branch of the twelfth, nerve confluent with the foramen lacerum posterius, or eliminated.
8. Sutural surfaces for the proötics abbreviated.

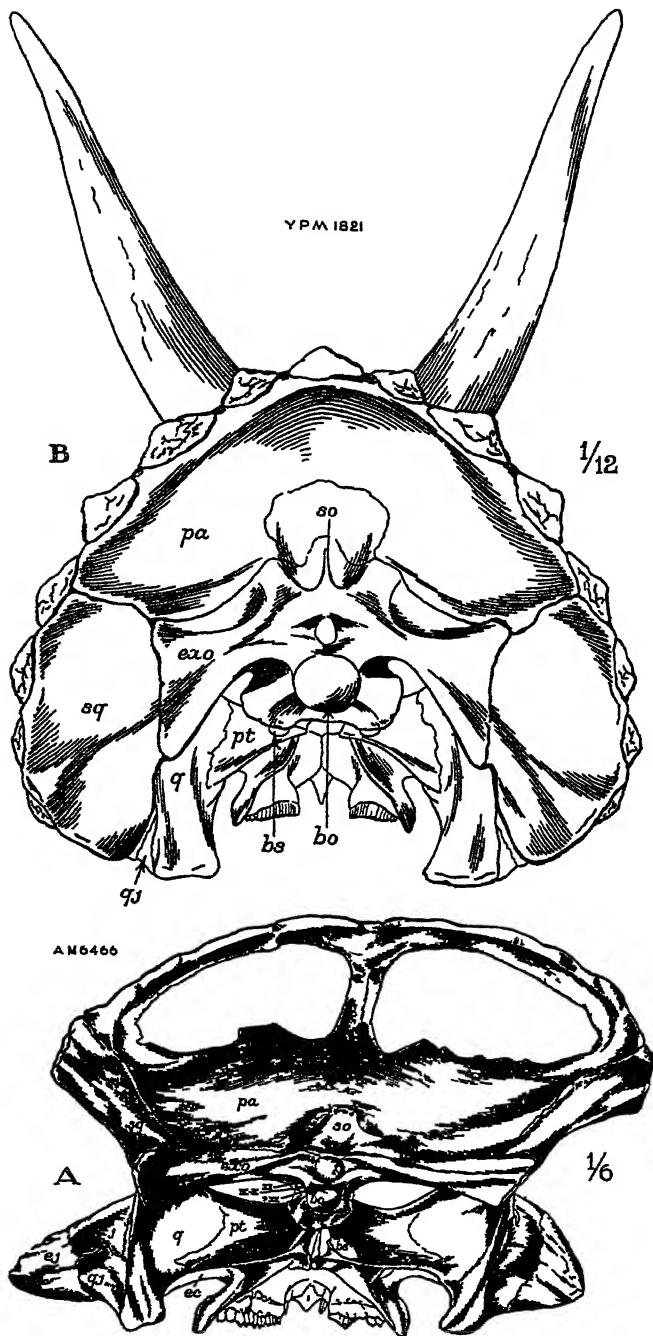


FIGURE 14. Posterior views of two ceratopsian skulls. A, *Protoceratops andrewsi*, fully adult. B, *Triceratops flabellatus*, modified from Hatcher, Marsh, and Lull.

Supraoccipital

The supraoccipital is proportionately larger, and more laterally extended than in any other known ceratopsian. It is also unique in that it forms the superior border of the foramen magnum. As pointed out earlier in this paper, in all other ceratopsians, the exoccipitals unite above the foramen magnum thus entirely eliminating the supraoccipital from the border of that foramen.

The lateral wing is pointed and much elongated. It intervenes between the parietal and exoccipital for about one-half the distance out to the distal margin of the latter. The depth is about one-third the width of the bone. The ventral border meets the exoccipitals in almost a straight transverse line, while the dorsal border is developed into a rather high median projection with a straight superior margin.

Anteriorly, the supraoccipital is well developed and occupies that part of the brain case which overlies the medulla oblongata and the postero-dorsal surface of the cerebellum. It is in contact with the prootics below, and has rather extensive contacts with the laterosphenoids in front.

Because of the lack of good material showing distinct sutures, the relationship of the supraoccipital to the laterosphenoids in the later ceratopsians is not always clearly understood. At least one species of *Triceratops*, however, sheds some light on this point. The left laterosphenoid (orbitosphenoid) of *T. eurycephalus* figured by Schlaikjer (1935: 59, fig. 4) is excellently preserved without distortion, and seems to show a distinct sutural surface for articulation with the supraoccipital just above the surface of contact with the prootic. This specimen shows, therefore, that in *Triceratops* the supraoccipital is still in contact with the laterosphenoid. Also, it shows definitely that in this genus the supraoccipital did not extend forward to above the foramen for the fourth nerve as described in *T. serratus* by Gilmore (1919: 109), and as shown in his figure 2 of the U. S. National Museum specimen (No. 2416), but that it extended forward only to above the exit for the fifth nerve as shown in Hay's figure of the same specimen (1909, pl. 2, fig. 1). As shown by Brown (1914: 547), the same condition prevails in *Anchiceratops*. In this genus also the laterosphenoid extends back to the exit for the fifth nerve.

In the later ceratopsians the supraoccipital becomes somewhat modified in form. Its lateral extensions become shortened and, posteriorly at least, it assumes the role of a keystone at the base of the enlarged frill. In spite of these modifications, the supraoccipital exemplifies the fact that regardless of the profound changes that

certain of the elements surrounding the brain have undergone in later ceratopsians, mainly as the result of the secondary roofing of the skull and the outgrowth of enormous brow horns, their associations with one another have remained essentially unchanged.

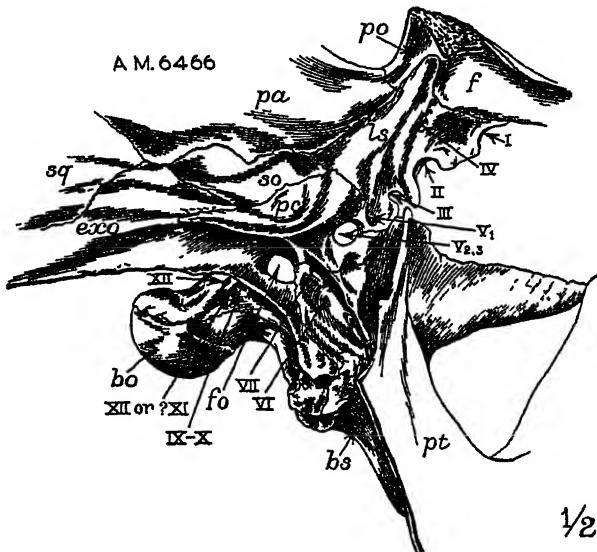


FIGURE 15. *Protoceratops andrewsi*. Outer side of the brain case of a fully adult "male."

Laterosphenoid

The laterosphenoid (= postoptic, 'orbitosphenoid, 'alisphenoid') is short and is antero-posteriorly expanded ventrally. The lower one-third has an almost vertical position. Above this the bone curves upward and abruptly outward so that the heavy and rather bar-like upper one-third, which is in contact with the frontal and postorbital, is nearly horizontal.

Approximately one-third of the ventral margin is in contact with the supraoccipital. In front of this, the laterosphenoid unites with the proötic to just above the foramen for the fifth nerve where it bends downward to the anterior margin of that foramen and extends downward to join the anterior portion of the basisphenoid (= probably the parasphe noid that has been completely fused with the basisphenoid). The distal part of the small, flat, narrow, posterior vertical projection of the pterygoid is braced against the side of this portion of the laterosphenoid. Just above this projection of the pterygoid there is a small

bar of bone that arches across the foramen for the ophthalmic branch of the fifth nerve. Both ends of this bar flatten and merge with the laterosphenoid without any indications of sutures. This bar, therefore, seems to be a part of the laterosphenoid, although the possibility of its being the epipterygoid should also be entertained.

The foramina in the laterosphenoid are distinctly shown in several skulls, notably in Am. Mus. No. 6466. (See FIGURE 15.) This bone forms a small portion of the anterior margin of the foramen for the fifth nerve, immediately in front of which is the bar of bone mentioned above that arches over a foramen leading from the foramen proöticum. This undoubtedly transmitted the ophthalmic branch of the fifth nerve. Immediately in front and below this foramen the laterosphenoid forms the superior border of the large foramen for the emission of the third nerve, which is bounded below by the parasphenoid. The front margin of the bone is broken so that only a slight portion of the border of the second nerve foramen is preserved above the foramen for the third. The foramen for the fourth nerve is rather large and is situated quite far above the foramen for the third. Near the top of the laterosphenoid, and in front of the median ridge are two foramina, close together. One is slightly in front of the other. These probably were exits for veins.

The main difference that exists between the position of the foramina in the laterosphenoid of *Protoceratops* and of *Triceratops* is that in *Triceratops* the exit for the fifth nerve is relatively more greatly separated from the exits for III and II. In this genus, the basal portion of this bone is proportionately much more expanded. The other striking differences in this element is that in *Triceratops* the upper portion is proportionately more extended, more erect, and much narrower and heavier.

Proötic

The proötic has virtually the same relationship with the other cranial elements as does that bone in the crocodile. Superiorly it is in extensive union with the supraoccipital, and posteriorly and posteroventrally it is well braced against the exoccipital, with its strong posterior projection reaching quite far out on the front of the exoccipital. At its postero-inferior corner it just touches the basioccipital. Ventrally it unites with the basisphenoid, and is in contact with the laterosphenoid throughout its anterior and antero-superior extent.

It forms nearly all of the border of the foramen proöticum, the superior border of the exit for the seventh nerve, and part of the anterior margin of the fenestra ovalis.

Basioccipital

As in *Triceratops*, the basioccipital forms no part of the foramen magnum border, although it is only barely excluded from that foramen in the young individual, in which the occipital condyle is made up almost entirely of the basioccipital. With age, however, the exoccipitals unite more extensively below the foramen magnum and form about one-third of the condyle in the adult individual. Also, with age, a more distinct "neck" is formed in front of the condyle. Anteriorly, the basioccipital is broadly expanded where it is firmly united with the basisphenoid.

In most of its characteristics the basioccipital of *Protoceratops* compares favorably with that of the later ceratopsians. It differs mainly in being proportionately longer, and in that it forms about two-thirds of the occipital condyle, whereas in the later members it constitutes only about one-third of that condyle.

Basisphenoid

In the young skull, the basisphenoid is proportionately large and elongated. With age, it is somewhat relatively shortened, although not as much so as in the later ceratopsians. It is much expanded posteriorly and seems to completely enclose the external opening of the middle eustachian canal which is located medially on the ventral surface just in front of the posterior margin. The basisphenoidal processes are relatively longer than in any other ceratopsian and are not received into as deep pits in the postero-ventral portions of the pterygoids. Neither are they as vertical in position as in the later forms.

On either side of the basisphenoid, fairly near the ventral margin, is a rather large foramen. These are probably carotid artery foramina. Anteriorly the basisphenoid presents a narrow and deep median septum, which extends upward to a level even with the ventral margin of the foramen proöticum where it is in contact with the laterosphenoid. It extends forward to, and is firmly buried between, the distal ends of the anterior ascending wings of the pterygoids where they unite postero-dorsally with the palatines. It is not in contact with the prevomers. There is no distinct suture between the anterior plate-like portion and the main body of the basisphenoid, although in skull Am. Mus. No. 6466 there is a suggestion of one. It is probable, however, that the entire portion anterior to the ventral point where the laterosphenoid and the proötic unite is the parasphenoid.

Pterygoid

The pterygoid is the most important element in the ventral architecture of the skull. It is the brace between the palatine, ectopterygoid, and maxillary in front, and the parasphenoid, basisphenoid, and quadrate behind. In this function it has assumed a very irregular form. In addition to the main body, which is braced between the basisphenoidal process and the palatine, there is a forked postero-lateral wing, a postero-dorsal wing, an antero-dorsal wing, and an antero-ventral wing. The main body is relatively limited in extent, and on the ventral surface near the inner margin there is a raised triangular area. From the front of this a ridge swings forward, and then downward where it continues along the posterior margin of the antero-ventral wing. Behind this ridge, the ventral surface is concave back to the margin that overhangs the basisphenoidal process—a concavity which extends back on to the posterior ventral surface of the postero-lateral wing. This concave area seems to be the beginning of what becomes a decidedly constricted groove in the later ceratopsians, to which Hatcher (Hatcher, Marsh & Lull 1907: 27) gave the name "eustachian canal."

The postero-lateral forked wing is the largest, and it overlaps considerably the posterior surface of the lateral wing of the quadrate. The lower fork is shorter than the upper fork and its distal end is not received in a well formed pocket or notch in the quadrate as in the later ceratopsians. The upper fork extends along the superior border of the quadrate almost to where the latter is nearly in contact with the exoccipital. The postero-dorsal wing branches off the postero-lateral at the basisphenoidal process. It continues up along the outside of this process and becomes very narrow near the extremity which terminates just below the opening for the ophthalmic branch of the fifth nerve where it is in contact with the laterosphenoid. The anterior margin of this wing is longer than the posterior margin. It is nearly vertical, becomes heavier ventrally, and forms the posterior border of a deep notch on the dorsal margin of the pterygoid. The anterior border of this notch is formed by the antero-dorsal wing which is high, narrow, and nearly vertical in position. The upper end of this wing curves abruptly forward and extends anteriorly for a third the length of the prevomer. This anterior extension is quite completely covered over by the prevomer. Mesially the antero-dorsal wing is suturally united with its mate except postero-dorsally where the anterior portion of the parasphenoid is tightly wedged between them. The front margin of this wing, together with that of the antero-ventral wing

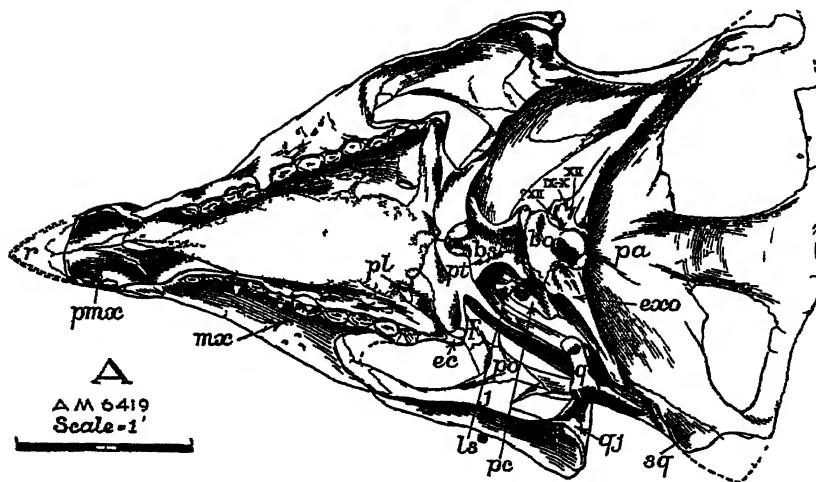
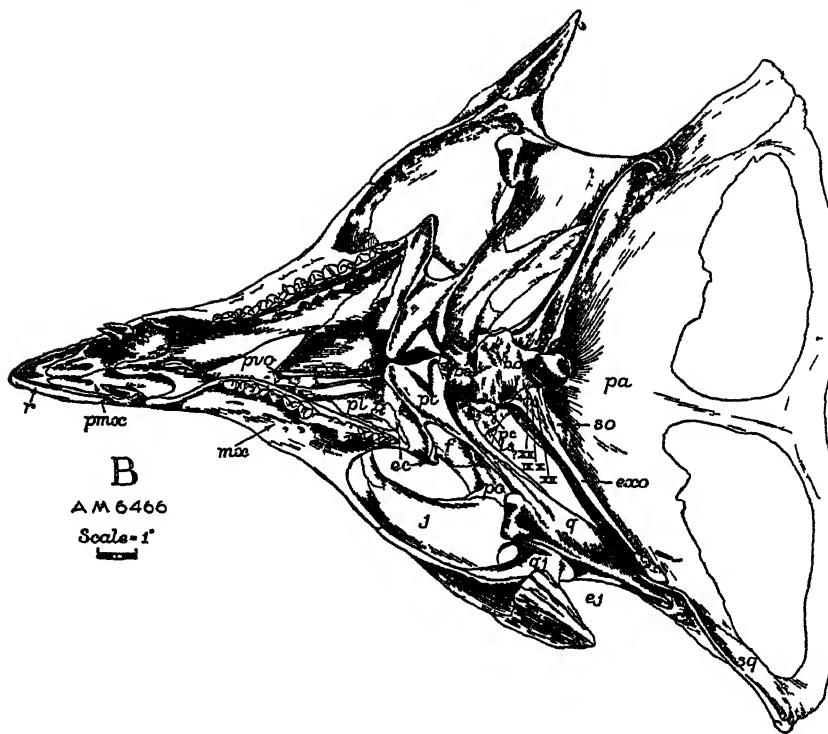


FIGURE 16. Two *Protoceratops andrewsi* skulls, ventral views, displaying the malar growth changes. A, very immature individual. B, fully adult individual.

gives an almost straight anterior margin to the pterygoid. Along a little more than the upper half of this margin, the pterygoid is in contact with the palatine. Below this the pterygoid in the very immature skull unites only with the ectopterygoid. In the older skulls, however, the alveolar margin of the maxillary extends back and unites with the pterygoid along the anterior margin for a very short distance. The narrow antero-ventral wing of the pterygoid projects quite far below this point. A rather large foramen is present between the ectopterygoid and palatine at the anterior margin of the pterygoid.

In the later ceratopsians the general features of the pterygoid are about the same as in *Protoceratops*. The main differences, as seen in *Triceratops*, are as follows:

1. The whole pterygoid is relatively shortened and the main body portion deepened.
2. A well formed "eustachian canal" is present.
3. There is no contact with the ectopterygoid on the palate, and contact with the maxillary is more extensive.
4. The postero-dorsal wing is reduced.
5. The antero-dorsal wing is in contact with its mate only at the uppermost portion.
6. This wing does not extend forward at its extremity and is not, therefore, embraced by the prevomer.

Ectopterygoid

The ectopterygoid is relatively larger in *Protoceratops* than in any other ceratopsian, and is also unique in that it is exposed on the palate. In all other known forms the maxillary is in contact with the pterygoid throughout the region below the palatine. Thus the ectopterygoid is eliminated from the palate and is a much reduced element situated on the postero-external alveolar portion of the maxillary and on the ventral antero-external surface of the pterygoid. In the young *Protoceratops* skull, as shown in Am. Mus. No. 6419, the ectopterygoid intervenes between the maxillary and the pterygoid (see FIGURE 16). With age, the alveolar portion of the maxillary grows backward across the ectopterygoid and unites with the pterygoid. The area of the ectopterygoid exposed on the palate is, therefore, proportionately reduced, thus foreshadowing the condition achieved in the later forms in which this bone is completely eliminated from the palate.

The greatest portion of the ectopterygoid constitutes most of the anterior margin of the subtemporal opening and forms a pulley-like ridge, from the alveolar portion of the maxillary to the jugal, across

which rode the pterygoideus anterior muscle. This ridge continues outward and upward onto the bar-like projection of the ectopterygoid that is braced over the maxillary-jugal suture beneath the orbit. Anteriorly the ectopterygoid forms the floor and the front end of the tunnel for the pterygoideus anterior muscle, and its antero-internal corner unites slightly with the internal postero-ventral projection of the lachrymal. Below this the front of the ectopterygoid is perforated by a rather large foramen for the maxillary branch of the trigeminus nerve.

Among the ceratopsians, the ectopterygoid has been recorded previously only in *Styracosaurus* (Lambe 1913: 112), and has been described only briefly in *Triceratops* (Hatcher, Marsh & Lull 1907: 26). Compared with the ectopterygoid of *Protoceratops*, this element in *Triceratops* shows the following main differences:

1. It has become proportionately very much reduced in size.
2. It has been eliminated from the palate by the posterior extension of the alveolar portion of the maxillary and the downward growth of the postero-inferior portion of the palatine.
3. The externo-dorsal bar has been reduced and it is, therefore, no longer in contact with the jugal.
4. It no longer forms a pulley-like ridge for the pterygoideus anterior muscle at the anterior border of the subtemporal opening.

Palatine

The palatine consists mainly of a high, thin plate that is expanded considerably antero-posteriorly, and that occupies a plane which is nearly parallel with the long axis of the skull. In addition, there is a narrow vertical plate which juts outward and backward from the front of this main plate. Dorsally it has a limited contact with the prefrontal, below which it is extensively in contact with the lachrymal. It tapers ventrally and merges with the main plate at the postero-inferior margin of the internal narial opening where the palatine sends forward a short projection along that margin.

The ventral margin of the palatine curves backward and downward where it is in contact with the maxillary, a contact which is scarcely greater than the contact with the ectopterygoid. Near its anterior margin, the latter is invaded by a ventral projection of the palatine which is weak in the young skulls and strongly developed in the adult. The increase in size of this projection from the young to the adult stage strongly suggests the manner in which the palatine, by its ventral growth, played an important role in eliminating the ectopterygoid from the palate in the later ceratopsians.

Throughout its posterior margin, the palatine is in contact with the pterygoid. Dorsally it fits firmly against the anterior projection of the latter, and continues forward onto the prevomer for more than one-third the length of that bone.

The palatine of *Protoceratops* has a general relationship to the other cranial elements that is usual for the ceratopsians. When compared, however, with the palatine of one of the later members of the group, such as *Triceratops*, it is seen to possess several primitive characteristics that are unique. These are as follows:

1. As a result of the extensive ectopterygoid exposed on the palate, the postero-ventral contact with the maxillary is much more limited.

2. There is no antero-inferior process that fits into a notch on the supero-internal surface of the maxillary. This is represented by a slight anterior projection above the maxillary at the postero-inferior margin of the internal narial opening.

3. There is no vacuity or "maxillopalatine foramen" such as Hatcher described in *Triceratops* that "passes from the cavity of the mouth to the infratemporal cavity" (Hatcher, Marsh & Lull 1907: 28). The presence of such in *Triceratops*, however, seems questionable.

4. The portion in contact with the pterygoid behind the internal narial opening is very much broader antero-posteriorly.*

5. The portion embracing the prevomer is very much more extensive.

Since the dorsal portions of the palatines are so extensively developed, the internal nares of *Protoceratops* occupy a more anterior position and are relatively smaller than in the later ceratopsians.

Prevomer

Together, the prevomers (= "vomers" of many authors) form a bar-like structure which serves as a beam between the antero-dorsal projections of the pterygoids that embrace the parasphenoid, and the anterior median portions of the maxillaries. They are suturally distinct even in fairly adult skulls, as is well shown in Am. Mus. No. 6408. Posteriorly they are deep and embrace, for about one-fourth their length, the anterior projections of the antero-dorsal wings of the pterygoids. They are in turn embraced by the palatines posteriorly

* In his figure of the skull of *Triceratops serratus* (Am. Mus. 970), Lull (1903: 688) shows the posterior borders of the internal nares formed by the pterygoids. Probably on the basis of this, he has suggested the same condition in *Monoclonius* (*Centrosaurus*) and *Chasmosaurus* (1933, figs. 5, 30). A re-examination of this specimen reveals that the posterior borders of the internal nares are formed by the palatines, and that these bones have about the same relation to the other elements as shown in the type of *Triceratops latifrons* as figured by Hatcher (1907, pl. 44).

for more than one-third their length. About half way forward, they become thin and in cross-section rounded. The anterior one-fourth, which is tightly wedged between the palatal portions of the maxillaries, becomes somewhat thicker.

In some respects the prevomers become quite modified in the later ceratopsians, insofar as can be judged by the only two specimens in which these elements have been recorded as completely preserved—a skull of *Chasmosaurus belli* (No. 2016) in the Yale Museum and a skull of *Triceratops serratus* (No. 970) in the American Museum. Compared with the prevomers of *Protoceratops* these elements in *Triceratops* show the following main differences.

1. The posterior ends are much expanded and are not as extensively



FIGURE 17. A portion of the left sclerotic ring of *Protoceratops andrewsi*.

embraced by the palatines; neither do they embrace as extensively the anterior tips of the antero-dorsal wings of the pterygoids.

2. According to Hatcher (Hatcher, Marsh & Lull 1907: 30) they are in contact with the anterior tip of the "alisphenoid" (= paraspheenoid). It appears, however, that in his figure 27 he has failed to show the upper ends of the pterygoids intervening between the paraspheenoid and the vomers. This character, therefore, seems doubtful.

3. They are triangular or dorso-ventrally flattened, and not oval or laterally compressed in cross-section as in *Protoceratops*.

4. They do not extend as far forward, and the anterior ends are much expanded. On the ventral surface, at least, they are not, therefore, wedged between the maxillaries.

In *Chasmosaurus*, according to Lull (1933, fig. 30) the prevomers form a more slender bar than in *Triceratops*. Also, they extend somewhat farther forward, the anterior ends are not expanded, and they are

wedged between the maxillaries for a short distance on the palatal surface.

Sclerotic Ring

Fragments of the sclerotic ring are preserved in several specimens. The most complete specimen, and the one which affords the most information, is Am. Mus. No. 6466—a fully adult individual. A nearly complete ring is preserved in the left orbit of this specimen. Some of the plates have been disturbed and it is difficult to determine their natural arrangement. Twelve complete plates and parts of three others are present, but the orientation of this preserved portion in the orbit is not certain. A clue to its orientation, however, is shown by the three overlapping plates in the upper right portion of FIGURE 17. Because of the direction of overlap of these plates, they represent either the dorso-posterior portion or the ventro-anterior portion of the ring. It cannot be determined, however, whether the first of these three is a plus plate or was overlapped by a plus plate which is not preserved. The third of these plates overlaps what apparently is a minus plate, and the two remnants of the fourth plate beyond this is probably the other plus plate. Accurate designation of the other plates is not possible.

THE ENDOCRANIAL CAST OF *PROTOCERATOPS* AND A COMPARISON OF IT WITH THAT OF LATER CERATOPSIANS

The cranial elements of a large "male" skull, Am. Mus. No. 6466, were most skillfully disarticulated and prepared by Mr. Otto Falkenbach, who was then able to obtain from this specimen an unusually perfect endocranial cast with the semicircular canals beautifully shown. (See PLATE 7.)

Endocranial casts have been recorded from only two other ceratopians, *Triceratops* and *Anchiceratops*. The endocranial cast of the latter, described and figured by Brown (1914: 545–548, pls. 34, 35), is as complete, however, as that of *Protoceratops*. It is possible, therefore, to make a detailed comparison of the casts of a primitive and an advanced, although in certain respects quite specialized, form. (See PLATES 7 and 8.) Endocranial casts have been recorded from three species of *Triceratops*. One of these, that of *Triceratops serratus* (U.S.N.M. 2065), is quite complete, although in none have casts of the semicircular canals been obtained. Marsh first described the endocranial cast of *T. serratus* (1896, pl. 77, fig. 4). His figure consists

of a lateral view of the cast, that was later reproduced by Hatcher, Marsh, and Lull in their monograph on the Ceratopsia (1907: 39, fig. 34), in which monograph was also published a ventral view of the same cast (p. 37, fig. 32). Hay (1909: 108) recognized the misidentification of certain nerves in these illustrations, and made the necessary corrections in his figures of a new, and much more detailed cast from the same specimen (pl. 3, figs. 1-3). He also published a lateral view of an endocranial cast from a specimen of *T. sulcatus* in the National Museum (pl. 3, fig. 4). The only addition that should be made to Hay's designation of nerves is that the nerve opening in the exoccipital which he regarded as emitting the tenth nerve probably was also the exit for IX and XI. In 1935, one of us (Schlaikjer: 60) published a figure of the anterior region of an endocranial cast of *Triceratops eurycephalus*.

When the endocranial cast of *Protoceratops* is compared with those of the later and more progressive ceratopsians, it is seen to possess quite a number of distinct features. This is best shown by the following comparison with *Anchiceratops*.

Protoceratops andrewsi

1. Endocranial cast proportionately large and deep.
2. Medulla oblongata* short and narrow. Constriction between it and cerebellum very marked.
3. Cerebellum and cerebrum together relatively long, and only a slight suggestion of processes on the cerebellum.
4. The optic nerves are close together and there is no bony partition between them.
5. III is behind the ventral portion of II.
6. IV is considerably above the dorsal margin of II.
7. VI behind, below, and close to V.
8. XI, or a separate branch of the hypoglossal, distinct.
9. Pituitary body relatively small and narrow, and directed downward.
10. Semicircular canals high, and the anterior one is considerably longer than the posterior.

Anchiceratops ornatus

1. Endocranial cast proportionately smaller and shallower.
2. Medulla oblongata somewhat longer and broader. Constriction between it and cerebellum less marked.
3. Cerebellum and cerebrum together relatively shorter, and large elongated processes on the cerebellum.†
4. The optic nerves are considerably separated and there is a bony partition between them.
5. III is behind, and below the ventral margin of II.
6. IV is behind, and below the dorsal margin of II.
7. VI below and quite removed from V.
8. XI, or a separate branch of the hypoglossal, not distinct.
9. Pituitary body relatively large and broad, and directed backward.
10. Semicircular canals low and the anterior and posterior ones subequal in length.

With the very marked increase in size of the later ceratopsian skull, in which there was special emphasis on the development of an enormous frill and very large horn-cores, a relatively smaller brain is to be expected in these later forms. Also that *Protoceratops* has only suggestions of the processes on the cerebellum is to be expected. Although on the endocranial cast these processes appear to be vertical extensions of the cerebellum, it seems certain that they represent nothing more than sinuses, which in some way seem to be correlated with strengthening of the region overlying the brain when the secondary roofing of the skull took place. The fact that the optic nerves in the later ceratopsians are considerably separated, and each is completely enclosed by the laterosphenoid, is directly related to the position of the orbit with relation to that of the brain case. In *Protoceratops*, the orbits have a more anterior position and the skull is relatively narrow between them. In the later forms, the skull becomes proportionately wider between the orbits, and the orbits shift posteriorly with respect to the brain case. The optic nerves, therefore, become more widely separated and more outwardly directed.

As stated above, under the discussion of the exoccipital, *Protoceratops* is unique in possessing a third foramen in the exoccipital. The endocranial cast shows that this foramen enters the brain cavity between the foramen for the hypoglossal nerve and the fenestra ovalis. What nerve passed through this foramen is not certain. One possibility is that it was the exit for the eleventh nerve. In this case, in the later ceratopsians, which do not have this separate nerve, the eleventh emerges with the ninth and tenth through the fenestra ovalis. On the other hand, since the eleventh is really nothing more than a branch of the tenth, it does not seem probable that it would first become separated, and later become confluent again with that nerve. Another possibility is that this foramen emitted a separate branch of the hypoglossal nerve, in which case the foramen would simply become eliminated in the later ceratopsians. This explanation is perhaps the more reasonable.

* While names of the various portions of the brain are used here, it should be understood, of course, that endocranial casts do not show the true dimensions and configurations of the brain, since the surrounding structures are not preserved, and the dimensions of the peridural and subdural spaces are not shown.

† Just in front of these processes in *Anchiceratops* the cerebral area is deeply concave. This is undoubtedly the result of crushing, as is indicated by the bone overlying this area. There is little reason for believing that the form of the dorsal surface of the cerebrum was peculiarly different from that in *Tyrannosaurus* which shows no such concavity.

SUMMARY OF THE SALIENT GROWTH CHANGES IN THE SKULL

As shown above, each of the forty-six skull elements undergoes some modification from the young to the adult stage. The changes reflected in certain elements, however, are greater than in others, and thus the adult skull possesses many features which are strikingly different from those present in the immature skull. The more salient of these growth changes are as follows:

Young

1. Nasal short, deep, broad and flat above, and only slightly notched anteriorly at the postero-superior border of the narial opening.
2. Frill short and only slightly expanded posteriorly. Median surface more or less on the same plane as the frontals, and median crest low. Temporal openings large and rounded in form.
3. No parieto-frontal depression.
4. Frontals large. Anterior projection elongated, exposure on superior border of orbit extensive, and dorsal contact with parietals broad.
5. Postorbital flat and smooth above. Superior projection pointed.
6. Orbita proportionately large.
7. Palate wide posteriorly.
8. Long axis of the narial opening anteriorly inclined.
9. Anterior branch of the premaxillary extends back to a position a short distance in front of the posterior margin of the narial opening.
10. Width of the anterior branch of the premaxillary rather narrow.
11. Post-alveolar extension of the maxillary not in contact with the pterygoid.

Adult

1. Nasal long, narrow above, and arched upward to form an incipient horn-core. Deeply notched anteriorly at the postero-superior border of the narial opening.
2. Frill elongated, greatly expanded posteriorly, and steeply inclined to the plane of the frontals. Median crest heavy and high. Temporal openings laterally constricted and elongated.
3. Well developed parieto-frontal depression showing the beginning of secondary roofing of the skull.
4. Frontals small. Anterior projection abbreviated and blunt, exposure on superior border of orbit restricted, and dorsal contact with parietals narrow.
5. Postorbital arched and very rugose above. Superior projection blunt.
6. Orbita proportionately smaller.
7. Palate proportionately wider posteriorly.
8. Long axis of the narial opening more erect.
9. Anterior branch of the premaxillary extends back to a position behind the posterior margin of the narial opening.
10. Width of the anterior branch of the premaxillary relatively broader.
11. Post-alveolar extension of the maxillary in contact with the pterygoid.

Young

12. Pocket at the inferior border of the preorbital fossa very pronounced.
13. Lachrymal large and the anterior border faces obliquely downward.
14. Prefrontal short, proportionately broad in front, and restricted to the antero-superior border of the orbit.
15. The anterior-posterior branch of the squamosal, short, shallow, and nearly horizontal.
16. Lateral temporal opening large, rounded above and in front, and the posterior margin quite erect.
17. Jugal rather horizontal in position, contact with the lachrymal limited, line of union with the maxillary oblique, area on orbital border quite extensive, depth under orbit shallow, and when seen from above it is in a rather vertical plane.
18. Area of ectopterygoid exposed on the palate large.
19. Exoccipitals do not enter into the formation of the condyle.

Adult

12. Pocket at the inferior border of the preorbital fossa reduced.
13. Lachrymal proportionately reduced and anterior border more erect.
14. Prefrontal elongated, narrow in front, and forms more than one-half of the superior border of the orbit.
15. The anterior-posterior branch of the squamosal elongated, deep and inclined upward anteriorly.
16. Lateral temporal opening somewhat reduced, dorsal border flattened, and the posterior margin steeply inclined.
17. Jugal rather vertical in position, contact with the lachrymal extensive, line of union with the maxillary quite erect, area of orbital border restricted, depth under orbit much increased, and when seen from above it flares out below bringing nearly all of the lateral surface into view.
18. Area of ectopterygoid exposed on the palate reduced.
19. Exoccipitals form approximately one-third of the condyle.

SUMMARY OF THE PRIMITIVE CHARACTERS OF THE SKULL

After an evaluation of the constant features of each element of the skull, and after a comparison of each part with that in the skulls of all the other ceratopsians, it becomes remarkably clear that *Protoceratops* is strikingly archaic in an extraordinarily large assemblage of characters. It should be understood, however, that these characters are determined as primitive relative to the net result of the evolutionary changes in the whole group, and that while the general evolutionary trends are constant in the later forms, an intermediate type may be as progressive, or even more so, in certain characteristics than an end member.

The more important primitive features, all of which are shown in the various illustrations above, are as follows:

1. Rostral occupies a rather high position on the front of the skull.
2. Narial opening small and no well-developed fossa in front of it.
3. Premaxillary proportionately large, and the top of the posterior branch even with, or above the dorsal margin of the lachrymal. Alveoli of two teeth present, and septum represented only by a welt inside the narial opening at the antero-inferior end where the premaxillaries meet.
4. Maxillaries deep and anterior margins steep. Alveolar portions not extended posteriorly, pre-alveolar space great.
5. Anterior ends of prevomers not expanded, and extensively wedged between the maxillaries on the palate.
6. Palate proportionately very wide posteriorly.
7. Preorbital fossa large.
8. Lachrymal large and the anterior border faces obliquely downward.
9. Nasal small and arched to form an incipient horn-core.
10. Orbit proportionately large.
11. Prefrontals do not meet in the midline to exclude the nasals from contact with the frontals, and they form part of orbital border.
12. Palpebrals freely articulate with the prefrontals, and are not, therefore, taken over into the skull roof to form the "supraorbitals." Not in contact with the lachrymals and postorbitals.
13. Postorbital mostly posterior to the orbit, and it presents only the first suggestion of a brow horn-core in that it is arched and rugose in the adult.
14. Lateral temporal opening large.
15. Frontals in contact with the nasals, form part of the orbital borders, and, together with the parietals, show only the beginning of the "frontal fontanelle" in the form of a parieto-frontal depression.
16. Frill short and has a high median crest. Entire structure acted as an anchor for the large capiti-mandibularis muscle masses.
17. Squamosal small, and with the postorbital, it forms the primitive narrow postorbital-squamosal bar which is nearly parallel with the horizontal plane of the skull. Does not unite with the jugal and (or) quadratojugal behind the lateral temporal opening.
18. Epoccipitals were not present.
19. Jugal short, broad, and quite oblique in position.
20. Epjugals large and rather pointed.
21. Quadratojugal with an extensive antero-inferior projection along the inside of the ventral margin of the jugal. The interno-ventral extension some distance above the articular surface of the quadrate.

22. The quadrate and the exoccipital do not unite—the posterior ventral projection of the squamosal intervenes. The lateral wing of the quadrate is quite clearly demarcated from the main shaft, much extended mesially, and on its infero-posterior surface there is no well-formed pocket or notch for reception of the overlapping process of the pterygoid.

23. Exoccipitals do not unite above the foramen magnum. Lateral extensions relatively long and slender, extend out to the lateral margins of the frill, and distal ends not expanded. Sutural surfaces for the prootics extensive.

24. Supraoccipital large, laterally extended, and it forms the superior border of the foramen magnum.

25. Laterosphenoid low and antero-posteriorly expanded ventrally.

26. Basioccipital proportionately long and forms about two-thirds of the condyle.

27. Pterygoid elongated and the main body is shallow. No well-formed "eustachian canal."

28. Ectopterygoid large, extensively exposed on the palate, is in contact with the jugal, and forms a pulley-like ridge for the pterygoideus anterior muscle at the anterior border of the subtemporal opening.

COMPARATIVE STUDY OF THE LOWER JAW

As in the case of the skull, the lower jaw is represented by a large quantity of material—a collection which includes numerous growth stages from the very youngest to the fully adult animal. One specimen, consisting of a left dentary, is unquestionably that of an unhatched individual.

Furthermore, as in the treatment of the skull, the procedure in the study of the lower jaw is to present a description of each element, and to make comparison of each with the same element in the later ceratopsians. Special emphasis is laid on the main growth changes, the primitive and variable characters, and on the main ontogenetic features that seem to prophesy certain evolutionary trends in the later ceratopsians. Summaries of these are given below.

The lower jaw presents no important characters which would eliminate *Protoceratops* as an ideal structural ancestor for the more progressive ceratopsians.

Predentary

The predentary is very sharply pointed anteriorly, is quite compressed, is rather shallow, and is considerably elongated. The dorsal

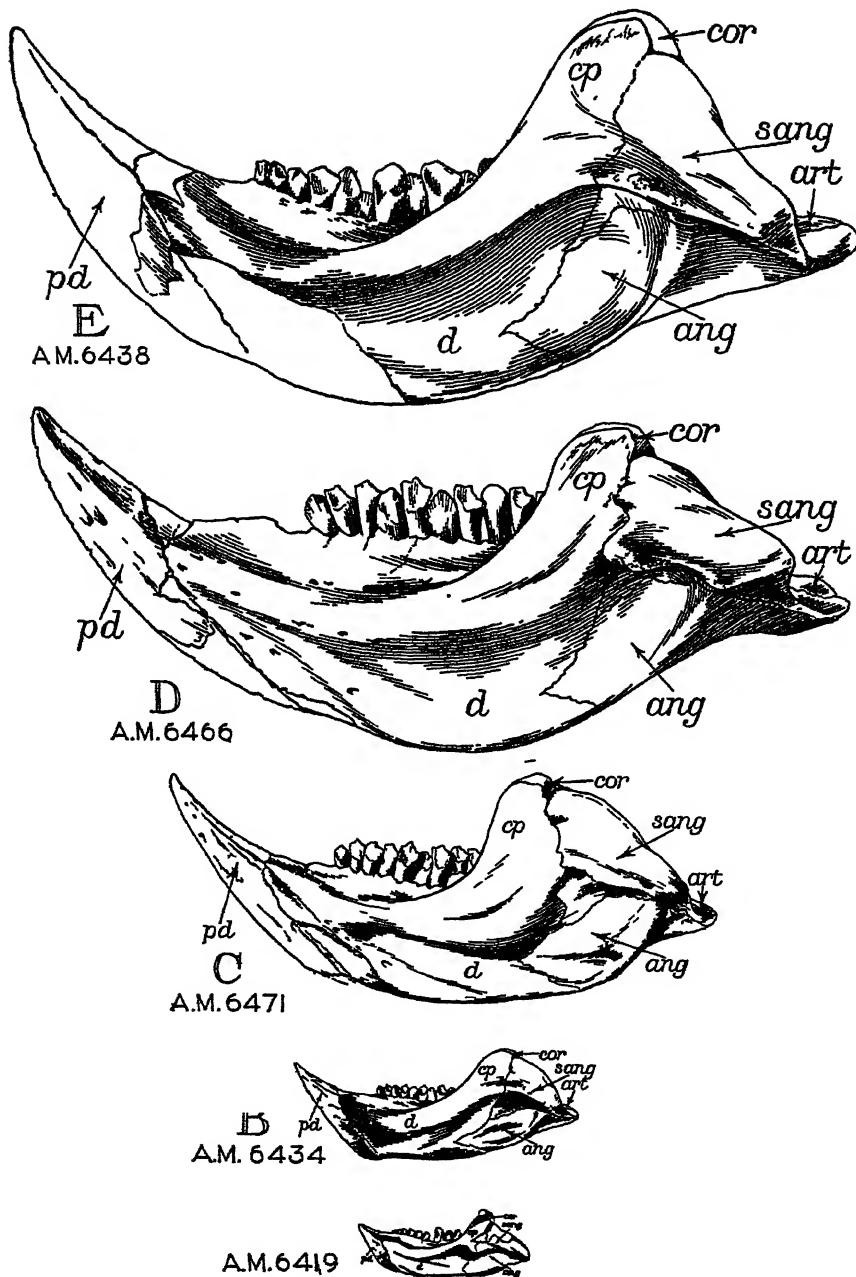


FIGURE 18. Series of *Protoceratops andrewsi* left lower jaws, external views, showing the development from a very immature to an old individual. One-third natural size.

margin is gently concave antero-posteriorly, and presents a broad flat surface which slopes outward. This margin is rather short, being less than half the length of the bone. Posteriorly it terminates in a short, blunt projection which slightly overlaps the superior margin of the dentary. The dorsal surface is slightly ridged medially where the inner edges of the margins come together. Behind this, it is shallowly concave and terminates medially with a short blunt projection that is wedged between the dentaries.

The ventral margin, in the adult, is gently convex antero-posteriorly and is transversely rounded and keel-like. There is a fairly marked constriction about mid-way down just in front of the upper part of the front of the dentary. The postero-ventral projection divides into two branches immediately in front of the splenials, which extend back a short distance on the ventral margins of the latter and the dentaries.

With age, several interesting changes take place in the predentary. In the very young individual the ventral margin is directed abruptly upward and is very straight throughout most of its length. Postero-ventrally, at a point located approximately below the posterior edge of the dorsal margin, it continues abruptly backward and downward so that in profile the margin forms an obtuse angle. With age, this angle is eliminated and the ventral margin becomes gently convex downward. Also, in the young jaw, as shown in Am. Mus. No. 6419, the dorsal margin is proportionately shorter. Likewise, with age the whole predentary becomes proportionately more elongated, broader, and more pointed. In the very young, however, the whole dorsal surface is quite rounded and the tip is very blunt—a primitive condition. These changes with age strongly suggest a foreshadowing of the principal features of the predentary in the later ceratopsians, for in a series including an adult *Protoceratops*, *Leptoceratops*, *Brachyceratops*, *Monoclonius*, and *Triceratops*, these very same changes are further emphasized. When compared with that of *Triceratops*, the predentary of *Protoceratops* is in several ways quite strikingly different. For example:

1. Posteriorly, the supero-lateral projections are far in front of the postero-ventral projection. This feature is also true of the dorso-median projection.
2. The dorsal margins are short and are not grooved. Instead they slope outward from the inner edges—an arrangement that probably was influenced by the presence of premaxillary teeth which overlapped the predentary posteriorly when the mouth was closed.

3. The premaxillary is much more pointed, its ventral margin is not as convex, and it is much less rugose. In addition it is proportionately shallower, longer, and narrower.

These are primitive characters, and the premaxillaries of *Leptoceratops*, *Brachyceratops*, and *Monoclonius* form an intermediate series showing how the specialized premaxillary of *Triceratops* was developed from this primitive type.

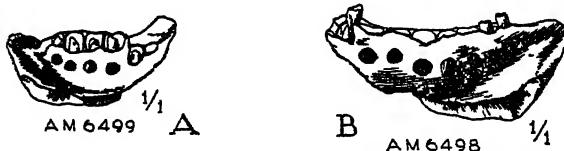


FIGURE 19. *Protoceratops andrewsi*. Inner views of two dentaries. A, from an unhatched individual. B, from a very immature individual.

Dentary

The series of *Protoceratops* dentaries is unusually complete including even one most unique specimen (Am. Mus. No. 6499) of an unhatched individual consisting of a right dentary. It is certain that this dentary is that of an unhatched individual. Five teeth, unworn, and each having the incipient form of a functional tooth, are deep-set in their alveoli. There is a low thin ridge of bone, not present in any individual with functional teeth, along the outer alveolar border that overhangs the alveoli somewhat, and bears only slight alveolar markings on its upper surface. The teeth, therefore, were not erupted. The estimated size of this individual based on the proportions of the beautifully preserved skull, jaws, and skeletal parts of a very young specimen, Am. Mus. No. 6419, could not have been larger than one of the medium-sized eggs found with *Protoceratops*.

The dentary is relatively shorter and deeper than in any other known ceratopsian. The anterior border is directed backward and downward and is overlapped considerably by the premaxillary. The symphyseal surface is extensive, but is interrupted ventrally where the front of the splenial intervenes and is lodged in a pit-like depression which is especially pronounced in very young individuals. The ventral border is nearly straight in the young jaw but with age it becomes sweepingly curved downward. This curving is reflected in the whole element and continues back onto the angular. It seems to be related to the relative increase in length of the premaxillary, the deepening of the

skull, and the more upright position of the frill which changes the direction of pull on the mandible in the adult and old skulls.

The coronoid process is low and narrow, and is set only a short distance out from the tooth row. The very tip is externally rugose, suggesting a muscle attachment. The upper part of its internal surface is covered by the coronoid. In front of this element all along the inner surface of the coronoid ridge is a well marked zone for the attachment of a portion of the capiti-mandibularis muscle mass that extends forward to about midway along the alveoli. In front of this, are several rather large foramina, probably for the emission of branches from the mandibularis branch of the trigeminus.

Postero-internally the dentary is overlapped by the large and deep splenial, and the intercoronoid extends along the inner alveolar margin for about half the length of the tooth-row. Down a short distance from the inner alveolar margin there are a series of foramina—one for each vertical series of teeth—which are connected by a shallow groove on the outer surface of the dentary.

The function of these foramina has been fully described in a previous paper (Brown and Schlaikjer, 1940b). Also, a comparison of the dentary of *Protoceratops* and the other ceratopsians has been made in that paper.

Angular

The angular is large and very irregular in form. Its ventral surface is antero-posteriorly convex downward. Anteriorly it is ridge-like, but posteriorly the ridge broadens under the articular and the articular portion of the surangular. Postero-internally it sends a heavy projection backward and upward under the articular and the internal projection of the surangular. The dorsal surface is deeply and openly grooved—thus forming the floor, and in part the sides of the posterior part of the Meckelian orifice. There is, therefore, an external and an internal dorsal lateral margin. The former is primarily in contact with the surangular, and extends far forward overlapping the dentary in the mandibular fossa. The latter is overlain throughout by the prearticular.

The whole external surface of the angular is deeply concave, especially in the adult, and the anterior portion of the internal surface is overlapped by the large splenial. This overlapping is not nearly so extensive, however, as in the later ceratopsians. In *Triceratops*, for example, as shown in the very excellently preserved left lower jaw of *T. sulcatus*, No. 4276, in the National Museum collection, the splenial extends back to the posterior margin of the jaw where it unites with

the articular. In front of this it covers the whole internal margin of the angular and the ventral margin of the prearticular. Other marked differences seen in the angular of the later forms are, the somewhat reduced size, the restriction of the portion exposed on the outer surface of the jaw, and elimination of the concavity of that portion. These changes seem to be correlated principally with the proportionate increase in size of the dentary and the articular.

Surangular

The surangular has the same general features, and relationship with the other jaw elements as in the later ceratopsians. In the following respects, however, it is quite distinct. First, the anterior portion is markedly deeper. In the later forms when the coronoid process of the dentary becomes distinctly differentiated, a blunt process of the dentary grows posteriorly for some distance between the angular and the surangular. The latter, therefore, becomes restricted in depth. Second, on the outer surface a heavy horizontal ridge is present just above the inferior margin. This ridge overhangs a concave zone along the inferior margin of the surangular which continues onto the angular below. It may mark the lateral zone of insertion of the pterygoideus posterioris. Posteriorly this ridge is flattened and it is continuous with the broad area of the ventral surfaces of the angular, which continues under the surangular. This probably marks the zone of attachment of the depressor mandibulae. Third, the postero-internal projection extends much farther around under and behind the articular. Fourth, the surangular forms about one-half of the surface for articulation with the quadrate, whereas in the later forms the surangular bears only a small portion of that surface. Also, the articular surface is extensive antero-posteriorly, and is much larger than the area of the surface of articulation on the quadrate. This shows that there must have been considerable anterior-posterior movement—certainly more than in the later forms.

Articular

The articular is relatively small and is completely embraced anteriorly, externally, and posteriorly by the surangular, and internally by the prearticular. The latter, however, extends up to the articular margin only where it unites with the surangular in front. Only the outer anterior area of the dorsal surface is slightly convex. Mesially to this there is a large concave area which articulates with the inner convex articular surface of the quadrate. The convex area extends

backward and inward forming a low ridge about where the postero-internal projections of the surangular and articular meet. Internal to this is another, but more limited, slightly concave area. The oblique ridge and this second concave area on the articular indicate that there certainly was an interior, and slightly outward movement to the lower jaw when the mouth was open.

In the later ceratopsians, the main changes in the articular seem to be, the proportionate increase in size resulting mainly in its forming more of the articular surface for the quadrate. Also, the postero-internal projection is reduced, and the projection of the surangular extending behind it becomes abbreviated so that the articular forms the most posterior limit of the mandible.

Prearticular

The prearticular is one of the smallest elements in the *Protoceratops* mandible. Posteriorly it unites with the antero-internal surface of the articular and remains at least partially unfused with the latter even in the old individuals. Ventrally it is in contact with the angular for about one-half its length, and antero-internally unites extensively with the splenial. Near the anterior margin, it forms the dorsal border of the internal mandibular foramen. Anteriorly it is below, and definitely is not in contact with the intercoronoid. Also, in all of the specimens in which the anterior margin of the prearticular seems to be completely preserved, it does not quite reach the dentary. These features are especially well shown in specimen Am. Mus. No. 6471. In the later ceratopsians when the alveolar portion of the dentary becomes extended posteriorly the intercoronoid is carried back considerably over the prearticular which is united extensively with the dentary. This is particularly well shown in the left mandible of *Triceratops sulcatus*, No. 4276, in the National Museum, when the splenial is removed.

A more extensive treatment of the relationships of the prearticular to the other jaw elements in the later ceratopsians has been given by us in a previous paper (1940b).

Splenial

The splenial is proportionately short and very deep posteriorly. Anteriorly it becomes abruptly shallow and meets the splenial of the opposite side at the symphysis. It extends down to, or nearly to, the ventral margin of the mandible throughout most of its length. The postero-dorsal projection unites above for a short distance with the

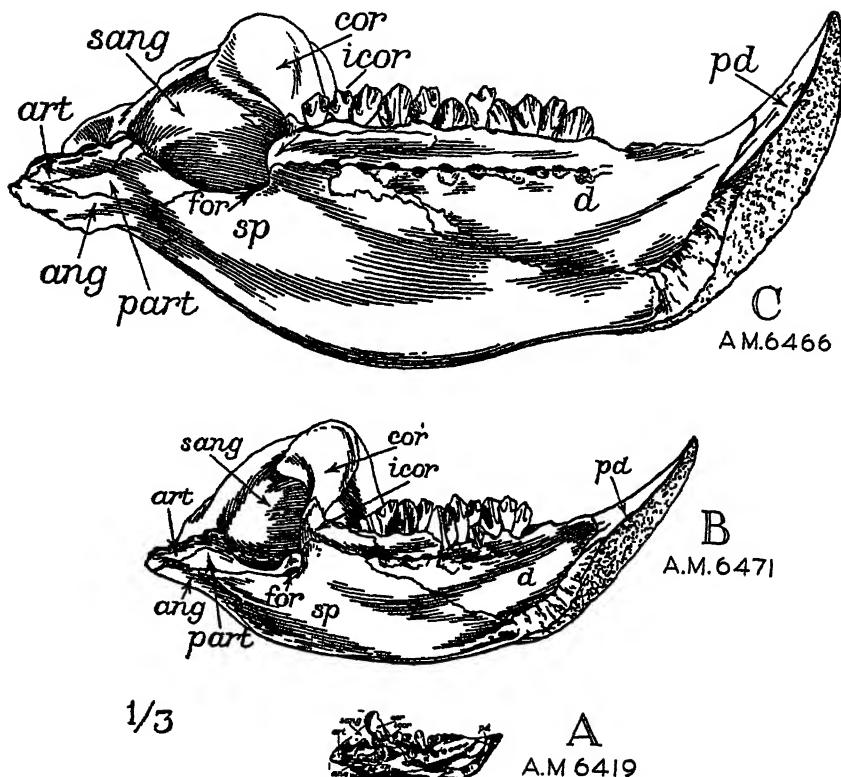


FIGURE 20. Three *Protoceratops andrewsi*; lower jaws, inner views, illustrating the changes from the very immature (A) to the fully adult (C) individual.

intercoronoid and with the latter forms a covering over the last few foramina on the inner side of the dentary. On the back slope of this projection there is a shallow notch that forms the interior margin of the internal mandibular foramen. Posteriorly it covers over only the anterior portions of the prearticular and the angular.

In its form and position the splenial of *Protoceratops* is primitive. In the more progressive ceratopsians this element becomes considerably modified, particularly in the following features:

1. As the alveolar portion of the dentary extends back and increases in depth, the splenial becomes shallow posteriorly.
2. With the above mentioned modification of the dentary, the postero-dorsal projection is carried downward and backward so that the internal mandibular foramen is almost entirely enclosed by the splenial.

3. Posteriorly it extends back and unites with the articular, thus entirely eliminating the angular and the ventral portion of the pre-articular from the lingual surface of the mandible. This change met the demand for greater strengthening in this region of the mandible when it assumed a more horizontal position instead of the upwardly curved form as in *Protoceratops*. Apparently, this straightening of the posterior part of the mandible was controlled by the more vertical pull of the jaw muscles, as was also, the high, clearly demarcated, erect coronoid process of the dentary.

Coronoid

The coronoid is relatively very large. The upper portion is very much expanded antero-posteriorly. Superiorly the anterior margin extends to, or nearly to, the front of the coronoid process, and the posterior margin extends back behind that process and overlaps the inner surface of the surangular. The posterior portions of the coronoid and the coronoid process of the dentary, therefore, form a pocket into which the upper portion of the front of the surangular fits securely. The superior margin projects slightly above the coronoid process, and shows markings, as does the whole of the inner surface of the expanded portion, of the capiti-mandibularis muscle insertion. Ventrally the coronoid narrows abruptly where it curves inward to meet the intercoronoid behind, and just outside the last alveolus. All specimens show that this element did not become fused with the dentary, although occasionally in a fully adult individual there is a tendency for it to become fused with the intercoronoid.

Three main changes take place in the coronoid of the later ceratopians. There is a proportionate reduction in size; the ventral portion is deflected backwards; and, the whole element shifts posteriorly so that in *Triceratops* the anterior margin scarcely reaches as far forward as the mid-line of the coronoid process, and the superior margin no longer projects above that process. These changes are interrelated with the posterior extent of the alveolar portion of the dentary and the clear demarcation, the antero-posterior expansion of the summit, and the erect position, of the coronoid process.

Intercoronoid

This is the smallest element in the mandible and, although not recorded in any of the ceratopians until recently (Brown and Schlaikjer 1940b), it is well preserved in many of the *Protoceratops* specimens, and is now known in several of the later forms. It is a thin, narrow

element that unites with the ventral projection of the coronoid, extends around back of the last alveolus, and continues forward along the inner alveolar border for one-half the length of the tooth-row. Postero-internally it unites with the dorsal projection of the splenial, and a ventral spur extends down behind this projection and nearly meets the prearticular, which it does in the later forms. This ventral spur is what becomes the floor of the intercoronoid in the later ceratopsians when this element is dragged backward as the alveolar portion of the dentary grows posteriorly.

The features of the intercoronoid in the later ceratopsians have been elaborated upon previously by us (1940b).

SUMMARY OF THE SALIENT GROWTH CHANGES IN THE LOWER JAW

The elements of the lower jaw, as in the case of those of the skull, all display some modification from the young to the adult stage. Certain of the elements, however, undergo greater transformation than others, which results in the development of rather marked changes in the mandible. (See FIGURES 18 and 20.) The three most important of these are as follows:

Young

1. Predentary short, narrow, and blunt. Most of ventral margin straight and directed abruptly upward. Dorsal margin short.
2. External surface of angular and lower part of surangular with slightly concave area, above which only a slight ridge is developed on the surangular.
3. Ventral margin of dentary straight, and posterior part of the mandible only slightly upturned. Surface for articulation with the quadrate below level of the tooth row.

Adult

1. Predentary more elongated, broader, and more pointed. Ventral margin convex downward. Dorsal margin longer.
2. External surface of angular and lower part of surangular with deeply concave area, above which is a heavy ridge developed on the surangular.
3. Entire ventral margin of the mandible sweepingly curved downward. Surface for articulation with the quadrate even with or above the level of the tooth row.

SUMMARY OF THE PRIMITIVE CHARACTERS OF THE LOWER JAW

When all of the seventeen elements of the lower jaws are analyzed individually, or as a complex, they confirm the evidence so admirably displayed in the skull, that *Protoceratops* is a remarkably primitive ceratopsian.

The more outstanding primitive features of the mandible are as follows (see FIGURES 18-20):

1. The predentary is shallow, long, and narrow. Dorsal margin short, not grooved, and its surface slopes outward from the inner margin.
2. Dentary short and deep, and ventral border curved downward. Coronoid process low, set close to tooth-row, and dorsal end not expanded. Alveolar area short, with few alveoli, and posteriorly it extends back to a position nearly opposite the middle of the coronoid process.
3. Angular large. External surface deeply concave and only part of the internal surface is covered over by the splenial.
4. Surangular deep anteriorly, has horizontal ridge on the external surface, forms approximately one-half of the surface for articulation with the quadrate, and postero-internal projection extends far around behind the articular.
5. Proportionately small articular.
6. Prearticular small, and not in contact with either the intercoronoid or dentary.
7. Splenial short, and very deep posteriorly. It forms only the anterior border of the internal mandibular foramen, it does not unite with the articular, and it covers only the anterior portions of the prearticular and angular.
8. Coronoid large, ventral portion not deflected posteriorly, and the dorsal portion is much expanded antero-posteriorly—covering most of the inner surface of the coronoid process of the dentary.
9. Intercoronoid not in contact with the prearticular.

COMPARATIVE STUDY OF THE TEETH

The dentition of *Protoceratops* is completely known in a large number of skulls and lower jaws representing all the important growth stages. The number of vertical series of teeth is variable. In the smallest hatched specimen, consisting of a left dentary (Am. Mus. No. 6498) approximately one and a half inches long, the stubs of three teeth and five alveoli without teeth show that the number was eight. A right dentary of an unhatched individual (Am. Mus. No. 6499), however, shows five unerupted teeth and a chamber for a sixth. (See FIGURE 19.) It would seem, therefore, that the minimum number for the hatched individual was probably eight. In the early adult skulls the number is from twelve to fourteen, and in the old skulls fourteen or fifteen. In two specimens, Am. Mus. No. 6417 and Am.

Mus. No. 6438, which are the oldest, and largest individuals represented in the collection, the number is fifteen. There seems to be no difference in the number of teeth between male and female skulls of relatively the same age. There are never more teeth in the mandible than in the maxillary and usually there are one or two more in the maxillary, but in the old individuals the number is fifteen both in the upper and lower jaw.

Protoceratops has fewer vertical series of teeth than any other known ceratopsian except *Leptoceratops*, in which form the exact number is unknown, but from the preserved alveoli in the dentary of that genus it would seem to have fifteen or fewer. In the later forms there always seems to be a greater number in the upper than in the lower jaw, and in *Triceratops* the discrepancy may be as great as ten or more.

As in all ceratopsians the dentition of *Protoceratops* is of the shear type. On the wear surfaces enamel is present only at the very margins of the tips, whereas in the later forms, such as *Monoclonius* and *Triceratops*, enamel covers a considerable amount of the whole tip on the wear surface of each tooth. On the opposite sides of the teeth—that is, on the inner surfaces of the lower teeth and the outer surfaces of the upper teeth—there is the customary thick layer of enamel. These surfaces are heavily ribbed, and each tooth has a single prominent vertical ridge, which is anterior to the mid-line on the lower teeth and posterior to the mid-line on the upper teeth. In the later ceratopsians the tendency is for this ribbing to become reduced and the vertical ridge to become heavier, higher, and more centrally located on the crown.

As in *Leptoceratops*, the teeth are single-rooted. A longitudinal groove is present on the anterior and on the posterior surfaces of each root. Into these grooves fit marginal portions of the crowns of the adjacent teeth, giving the root somewhat of an hour-glass form in cross-section. The next succeeding tooth is located at the tip of the root. Each tooth erupts, therefore, by the united effort of three teeth—one in front of, one behind, and one at the tip of the root. In the later ceratopsians, the root of each tooth becomes transversely bifurcate by the root dividing and spreading over the crown of the next succeeding tooth. The crown of the succeeding tooth is not, therefore, as far removed from that of the functional tooth. This arrangement in the later forms means, in the first place, a more even wear surface to the whole tooth battery. In *Protoceratops* with the succeeding tooth so far removed from the crown of the functional

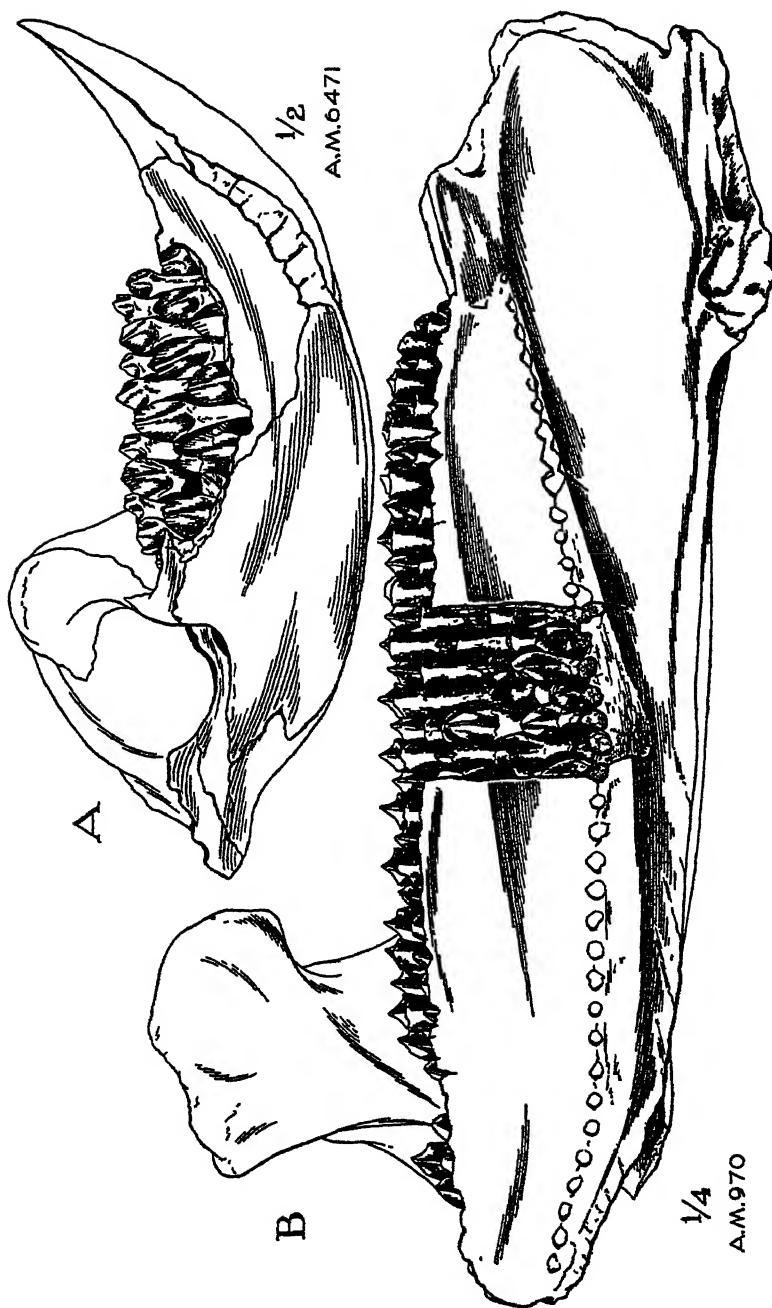


FIGURE 21. Inner views of two ceratopian left lower jaws. A, *Protoceratops andrewsi*. B, *Triceratops serratus*.

tooth, because of the long root of the latter, a considerable gap is present at the functional end of every other vertical series, although since the crowns are antero-posteriorly fan-shaped, and each is overlapped slightly by the one behind, this space is somewhat reduced by the adjacent teeth until the new tooth erupts. (See FIGURE 21A.) In the second place, in the tooth arrangement of the advanced forms there is far less chance for anterior or posterior displacement of any vertical series because the heavy crest-like median ridge of the crown becomes buried in the bifid root of the tooth which it is succeeding. Moreover, such displacement is entirely prevented in these forms by the appearance of a most unique structure consisting of a tube-like casing for each vertical series of teeth which is made not of cementum, but of spongy bone. This structure is formed at the base of each vertical series and erupts, and is worn away with the teeth. (See FIGURE 21B.) There is no such structure around the teeth of *Protoceratops*. In the evolution of the ceratopsians, the appearance of this unusual structure seems correlated with three main changes in the dentition—the development of two-rooted teeth, the increase in number of teeth in each vertical series, and the change from what is virtually a vertical series of only two, or possibly at the most three, teeth in each series to a pronouncedly curved series of as many as five or more teeth.

The arrangement of the whole battery of teeth is typically ceratopsian. Anteriorly and posteriorly the tooth magazine is shallower and the teeth are smaller than in the central area. In the manner of wear, however, the dentition of *Protoceratops* is different. The teeth of either alternate group present greater wear posteriorly in the jaw. That is, in the front of the battery a tooth may show only the first stages of wear, and proceeding posteriorly each homologous tooth in every other vertical series presents greater and greater wear so that towards the back of the battery, the homologous tooth may be quite completely worn away. In other words, the whole magazine of teeth is so arranged in the lower jaw that the teeth erupt upward and slightly forward, and in the upper jaw downward and slightly forward, which simply means that the teeth in the back portion of the jaw always wear out before those in the front portion. In the later forms, such as *Triceratops*, the condition is just the reverse. The lower teeth erupt upward and slightly backward, and the upper teeth downward and backward. Thus, the teeth in the front portion of the jaw wear out first. This change in wear seems correlated with a change in form and shear action in the lower jaw.

As pointed out earlier in this paper, the lower jaw of the adult *Protoceratops* is decidedly curved, the surface for articulation with the quadrate is even with or above the level of the tooth row, and the pull of the capiti-mandibularis mass is oblique. The lower jaw, therefore, had a pruning-shears type of action. As the jaw closed, each tooth in the lower jaw moved up and across one or more in the upper jaw. Thus to bring about the most efficient shearing with this type of action, the whole magazine of teeth erupted upward and slightly forward in the lower jaw, and downward and slightly forward in the upper. This results in the teeth of the posterior area of the dentition disappearing before their homologues in the anterior area. In the later ceratopsians, as in *Triceratops*, the lower jaw is straight, the surface for articulation with the quadrate is below the level of the tooth row, and the pull of the capiti-mandibularis mass is quite vertical,—a combination of characters adapted for a chopping action. Under these conditions, for the teeth in the lower jaw to shear upward and across those in the upper jaw, thus giving the most efficient type of shear, the whole magazine of teeth has to be arranged in the lower jaw so that the teeth will erupt upward and slightly backward, and in the upper jaw downward and slightly backward. Thus the teeth in the anterior part of the battery will disappear before their homologues in the posterior section.

MEASUREMENTS OF THE SKULL AND LOWER JAW

In the development from the young to the old individual, not a single important feature remains constant. To show diagrammatically the change of one feature with respect to any one other feature would

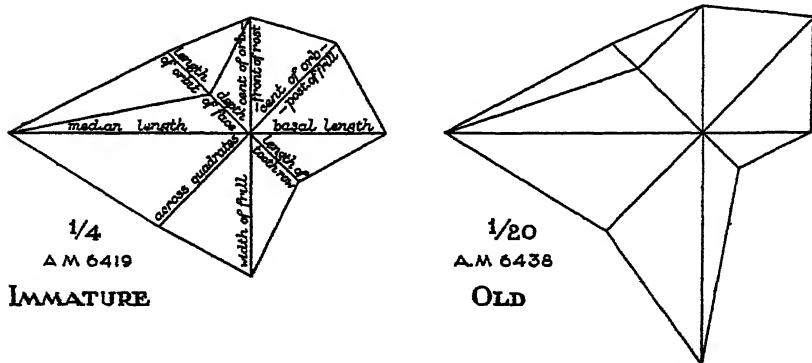


FIGURE 22. Diagrams illustrating how the proportionate changes that take place in nine dominant features of the skull from the young to the old growth stage are interrelated.

TABLE I
MEASUREMENTS OF THE SKULL

Am. Mus. No.	Length of orbit	Center of orbit to front of rostral	Center of orbit to posterior end of crest	Median length of skull	Greatest width of frill	Greatest depth of face	Basal length of skull. Condyle to anterior end of maxilla	Width across quadrate	Parieto- frontal suture to back of frill	Parieto- frontal suture to beak	Length tooth row on alveoli
6419	32.5	62	67	130	76	30	73	70	48.5	80	36.5
6408	58	180	202	369	251	117	202	185	165	213	96
6414	80	294	325	566	503	191	300	313	240	330	
6466	93	300.5	332.5	611	475	174	304	298	260	355	138
6438	92	341	403	692	622*	248	291	363*	310*	382	130
6425	81	277.5	357.5	600	473	189	262	266	270	322	
6439	63	284.5	271.5	470	435	168	235	235	205	268	111
6429	76	232	270	482	343	140	229	285*	203	281	110
6433	70	225	235	450	320	137	235	171*	191	258	99
6471	56	192	246	406	253*	126	183	125*	183	125	36

* Estimated

not, therefore, give a true picture of the real proportionate changes that take place. It is desirable, to have some method by which the changes of several dominant features can be diagrammatically shown in relation to one another. Such a presentation is given in FIGURE 22. Inspection of these diagrams will reveal at once how such proportionate changes as frill widening and lengthening, orbit reduction, and facial deepening are interrelated. The measurement data (in millimeters) used in the construction of these figures is given in TABLES 1 and 2.

TABLE 2
MEASUREMENTS OF THE LOWER JAW

Am. Mus. No.	Depth of jaw in front of coronoid ridge	Total length	Length of dentary. Antero- superior end to posterior	Length of tooth row	Height of coronoid above aveolar margin	Total length of pre- dentary	Length of dorsal surface of pre- dentary	Depth of jaw under coro- noid
6419	18	74	47	34	7	23	10	24
6408	50	225	120	93	40	112	54	90
6414	86	346	196	—	—	177	81	—
6466	84	361	202	142	52	165	82	128
6438	87	360*	190	145	74*	—	—	146
6425	72	327*	190	124	66	150*	59*	128
6439	67	300	162	112	47	136	70	110
6429	76	—	169	106	57	—	—	110
6471	51	236	125	86	46	122	57	90

* Estimated.

COMPARATIVE STUDY OF THE VERTEBRAL COLUMN

It is a remarkable fact that, insofar as can be determined, the number of presacral vertebrae is constant in all known ceratopsians. As shown elsewhere by the senior author (1917: 288), the most reliable criterion for distinguishing the change from the cervical to the dorsal series is the rise of the capitular facet from the centrum to the neural arch. Accepting this, the number of cervicals is always nine, and the centra of the first three are always fused. The number of dorsals is twelve. As will be shown later, the number of sacrals coossified by their centra is somewhat variable in the group. *Protoceratops* has eight in the fully adult. The number of caudal vertebrae in this genus, however, is unknown. The most completely preserved caudal series is shown in specimen Am. Mus. No. 6417, which has thirty-

two articulated and completely preserved vertebrae. On the last of these, the neural spine is still quite long, thus indicating that its position was rather far from the end of the series. The total number of caudals was probably over forty.

Cervicals

At least in one specimen (Am. Mus. No. 6418) a separate element is distinctly shown in front of the first cervical (see FIGURE 23). It consists of a semicircular band of bone, with a slight median postero-ventral projection, that forms the inferior margin of the cup on the anterior extremity of the atlas. It is almost completely fused with the latter, and is undoubtedly the homologue of what Lull described as the first cervical in the type of *Triceratops prorsus* (Hatcher, Marsh, and Lull 1907: 47), but what he later regarded as only a "sutural marking" on the atlas of this species and on that of *Centrosaurus*, and stated, "It has been suggested that the portion of the atlas in front of this suture may represent a proatlas, . . ." (1933: 40). That this element is homologous with the proatlas is not probable. The so-named "proatlas" is always located above the neural canal. It is composed of a pair of elements, which may become fused, and has the form of a vestigial neural arch. The most satisfactory explanation of its origin seems to be that it is formed by the dorsal arcualia (interdorsals) of the anterior half of the sclerotome which gives rise to the neural arch (formed from the basidorsals) of the atlas, and to the hypocentrum (formed from the basiventrals) that is sometimes present (as in the crocodile and, by homology, in numerous extinct reptiles) in front of the centrum of that vertebra. The interventrals of this anterior half-sclerotome either form the tip of the odontoid process, or are not developed. There is a possibility, therefore, that the element in front of the antero-ventral margin of the atlas of *Protoceratops* was formed from the interventrals of the anterior half-sclerotome, of which the interdorsals would normally give rise to the proatlas. The more probable explanation of its origin, however, is that it was derived from the basiventrals of the posterior half-sclerotome, of which the basidorsals gave rise to the neural arch of the atlas. In other words, it is nothing more than a hypocentrum. This is the condition in the crocodile, in which form there is not only a hypocentrum (intercentrum) in front of the centrum (pleurocentrum) of the atlas, but also one between the centra of the atlas and axis. In either case, the element in front of the atlas of *Protoceratops* cannot be homologous with what is known as the proatlas. That it is really a hypocentrum (intercentrum) seems most probable.

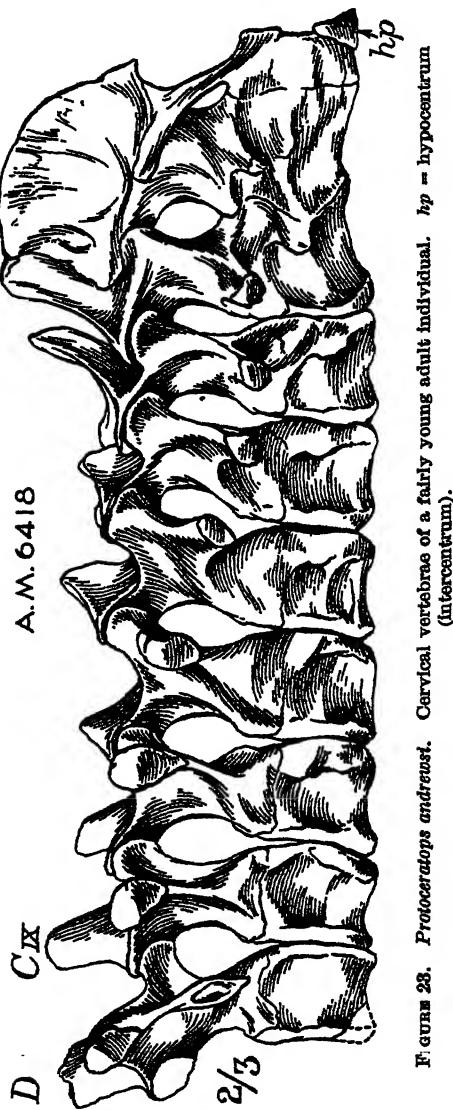


FIGURE 28. *Protoceratops andrewsi*. Cervical vertebrae of a fairly young adult individual. *hp* = hypoccentrum (intercentrum).

The centra of the first three cervical vertebrae are completely fused. The atlas is the smallest of the cervicals, and consists mainly of the centrum, the anterior surface of which is deeply concave for reception of the occipital condyle of the skull. The neural arch is represented by a narrow bar of bone on either side that extends upward and backward to the base of the neural spine of the axis. Its dorsal end flares

antero-posteriorly and is received into a shallow groove on the side of the arch of the axis. The posterior part of this antero-posterior expansion probably corresponds to the postzygapophysis of the atlas. At the base of this arch there is a low projection which is the capitular facet for a long, nearly holocephalous rib.

The centrum of the axis bears a dichocephalous rib although the tubercular facet is considerably smaller than that of any of the succeeding vertebrae. The intervertebral foramen between the axis and the third cervical is considerably larger than that foramen between the atlas and axis. The neural arch of the axis is larger than that of any of the cervicals. The neural spine is extremely expanded antero-posteriorly, being approximately three times larger than that of the third cervical. It extends posteriorly to a point above the middle of the centrum of the third cervical. By their rugosity, the dorsal and dorso-lateral borders show clearly the origin of the large rectus capitis posterior muscle mass. The origin of the obliquus capitis magnus is distinctly shown by a large, more or less circular depression near the base of the spine. At the posterior margin of the spine about half way down is a well-marked semicircular area which probably was the area of insertion of the spinalis and the semispinalis cervicis muscles. There was no interspinalis muscle between the axis and the third cervical since the spines and zygapophyses of these two vertebrae are closely applied to one another. They are not coössified, however.

Cervicals four to nine are all subequal in size. Of these, four and five are the most distinctive. Ventrally they are crest-like, while the ventral surfaces of six, seven, eight, and nine are rather broad and flat. In addition, the transverse processes of these two vertebrae are shorter and extend outward and abruptly downward. The transverse processes on the sixth are directed only slightly downward, on seven they are at right angles to the vertical axis of the vertebra, and on eight and nine they are directed slightly upward,—more so on nine than on eight. The capitular facet is most strongly developed on the fourth. On the succeeding cervicals it gradually diminishes, and occupies a higher position on the centrum. In specimen Am. Mus. No. 6418, which is a fairly young adult individual, the centra of the fourth and fifth cervicals are fused on the left side. This is undoubtedly a pathological condition which is not uncommon among the Ceratopsia. Such a condition has been reported in the fifth and sixth cervicals of *Monoclonius (Centrosaurus) flexus* by Lull (1933: 40) and in the fourth and fifth dorsals of *Styracosaurus parksii* by Brown and Schlaikjer (1937: 5).

When the cervical vertebrae of *Protoceratops* are compared with those of the later and more progressive ceratopsians, they are seen to be decidedly primitive. The following are the main changes that have occurred in the cervicals of the later forms:

1. The hypocentrum (intercentrum) in front of the atlas becomes ring-like, and the centrum of the atlas more deeply concave anteriorly.
2. The neural arches of the first three cervicals are lower, they become completely coössified, and the intervertebral foramina are reduced.
3. The capitular facet on the centrum of the atlas becomes reduced (*Monoclonius*) or is lost (*Triceratops*).
4. The neural spine of the axis loses its erect, hatchet-shaped form; becomes narrow, low, and more posteriorly directed; and, extends backward to a point above the posterior margin of the centrum of the third cervical.

Dorsals

As in all of the known ceratopsians, the number of dorsals is twelve. In the old individuals the centrum of the twelfth has a tendency to become slightly coössified with the centrum of the first sacral. At least this is the condition in specimen Am. Mus. No. 6466, which is the only fully adult specimen in which the sacrals and dorsals are known (see FIGURE 24).

On the first dorsal the capitular facet has shifted to the neural arch, and occupies a position on this and the second dorsal vertebra about midway between the centrum and the transverse process. On the remaining dorsals it moves up to the base of the transverse process where it remains throughout the series. In the later ceratopsians, however, it migrates outward on the ventral surface of the transverse process. In the posterior dorsals of *Triceratops* its position is midway between the apex and the base of that process. It is always oval in form and seems to be largest on the seventh dorsal.

The transverse processes are most robust on the first dorsal. They increase in length from one to three, and although four and five are as long as the third, they are slightly more slender. From six to twelve they decrease in length and robustness, and are markedly shorter on eleven and twelve. This is natural, however, since those two vertebrae are situated in the narrow space between the anterior ends of the ilia. They are directed upward and backward at about the same slant as the neural spines on dorsals one to five, but from six to twelve they are directed more outward and backward.

The neural spines increase in length from the first to the ninth

dorsal, and their antero-posterior diameter becomes greater from the first to the fifth, and is about the same for five, six, and seven. From the tenth to the twelfth the neural spines become shorter, and the antero-posterior diameter decreases from eight to twelve—especially so, on ten to twelve.

The intervertebral foramina decrease in size posteriorly. This results mainly from the change in form and position of the neural arches which expand antero-posteriorly and decrease in height posteriorly in the series.

The backward slant of the neural spines and arches is a prevalent feature in *Protoceratops*, and is most pronounced in the median area of the dorsal series. This feature is particularly characteristic in the young and early adult individuals, but is less dominant in the fully adult and old specimens. This probably is a primitive character, since in the later ceratopsians the tendency is for the neural spines and arches to become more erect and proportionately higher. There is, however, some variation in this respect in the later forms, as in *Monoclonus* (*Centrosaurus*), which may be due in part to age, and in part to individual variation. As is to be expected, the incisions of the neural arch under the prezygapophyses are shallower, and those under the postzygapophyses are deeper when the backward slant of the neural spine and arch is greater.

The centra are rather circular in outline when viewed from the front, and both the anterior and posterior ends are quite flat or somewhat shallowly concave. In diameter and length the centrum of each of the first three dorsals is greater than that of four to nine. It again enlarges from ten to twelve, and is largest on the twelfth. These characters of the centra are essentially the same as those in many of the later ceratopsians. In *Triceratops*, however, the tendency is for the ends of the centra to become pear-shaped in cross-section.

Sacrum

The number of sacrals in the immature individual is unknown. In specimen Am. Mus. No. 6418, which is a young adult, the number is seven, and of these, the sixth is incompletely coössified with the fifth and seventh. In the fully adult, as shown by Am. Mus. No. 6466, the first caudal (sacro-caudal) becomes coössified with the seventh, thus bringing the total number of sacral vertebrae, whose centra are coössified, to eight (see FIGURE 25).

In the young adult specimen (Am. Mus. No. 6418) the twelfth dorsal is completely distinct from the first sacral. In the fully adult

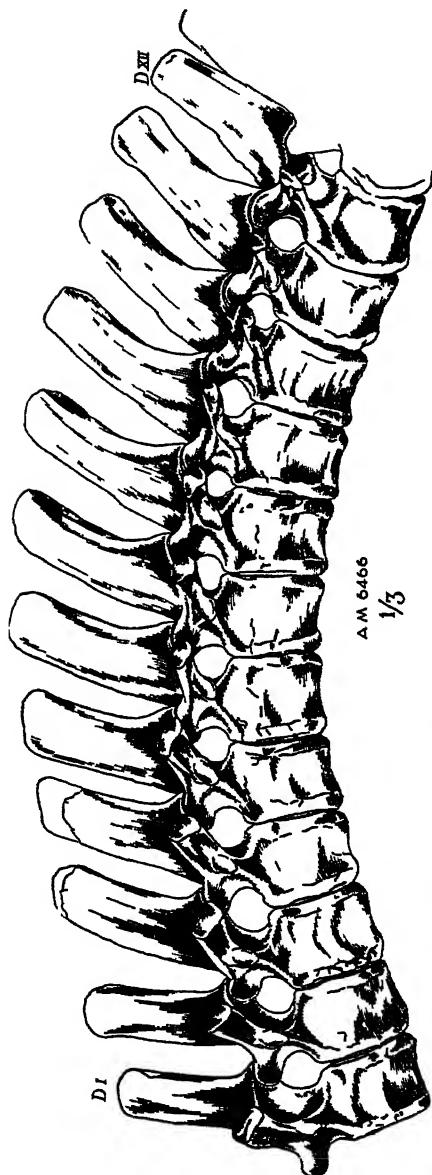


FIGURE 24. *Protoceratops andrewsi*. Dorsal vertebrae of a fully adult individual.

specimen, however (Am. Mus. No. 6466) the centra of these two vertebrae show slight coossification, and their neural spines are closely applied to one another throughout half their length.

The neural spines of sacrals one to seven are more or less of uniform

height, and although they are closely applied to one another, throughout most of their lengths, they are never coössified. The antero-posterior diameter of the spines is greatest on two to five. The zygapophyses are only partially coössified in the younger specimens, but in the fully adult they are quite completely fused. The prezygapophyses of the sacro-caudal, however, are not at all coössified with the postzygapophyses of the seventh sacral in the young adult specimen, and are only slightly so in the fully adult form.

The diapophysis of the first sacral is short, and bears a heavy rib which is about twice its length. This rib is directed forward, and its antero-posteriorly expanded distal end unites with the medio-ventral margin of the ilium just behind the anterior end of that element which is in contact with the rib of the last dorsal. Posteriorly the expanded distal end of the first sacral rib unites with the distal end of the diapophysis of the second sacral.

On the lateral posterior margin of the centrum of the first sacral, and extending two-thirds of the way back onto the side of the centrum of the second sacral, is a large facet or parapophysis. Similar but gradually smaller facets occur on sacrals three to five. Each facet is connected with the short, heavy diapophysis above by a strong low ridge. Articulating with, and partially or wholly coössified with, the parapophysis, or facet, the connecting ridge, and the under side of the diapophysis is a short, heavy, and highly modified rib.* Ventro-distally this rib is expanded antero-posteriorly, and together with the distally antero-posteriorly expanded diapophysis, it forms an I-beam structure. This structure is deepest on the second sacral—extending from the ventral margin of the pre-acetabular portion of the ilium down to the distal end of the pubic peduncle. It becomes less deep on sacrals three to five. The ventro-distal expanded ends of these ribs on sacrals two to five unite, as in all known ceratopsians, to form the acetabular bar. There is only a slight suggestion of this I-beam structure on the sixth sacral. On this vertebra presumably the short diapophysis and low parapophysis merge to form a very abbreviated, rather quadrangular projection, which is located partially on the

* These sacral ribs are sometimes referred to as parapophyses in the later ceratopsians (Hatcher, Marsh, and Lull, 1907: 51-53, Lull, 1933: 49). This is probably because the sacral ribs, in the adult form, become so thoroughly coössified with the centra. Nevertheless, in spite of the thorough coössification, the line of union is usually distinctly preserved, as is shown in the various specimens in the American Museum collection, and in Marsh's illustration (1896, plate 65) of the sacrum of an adult *Triceratops prorsus*. Furthermore, in the sacrum of *Agathomas sylvestris* (Am. Mus. No. 4000), which is obviously that of a young adult individual, the sacral ribs are suturally distinct, and were so described by Hatcher in the Ceratopia monograph (1907: 110). They should not be regarded as parapophyses in any ceratopian.

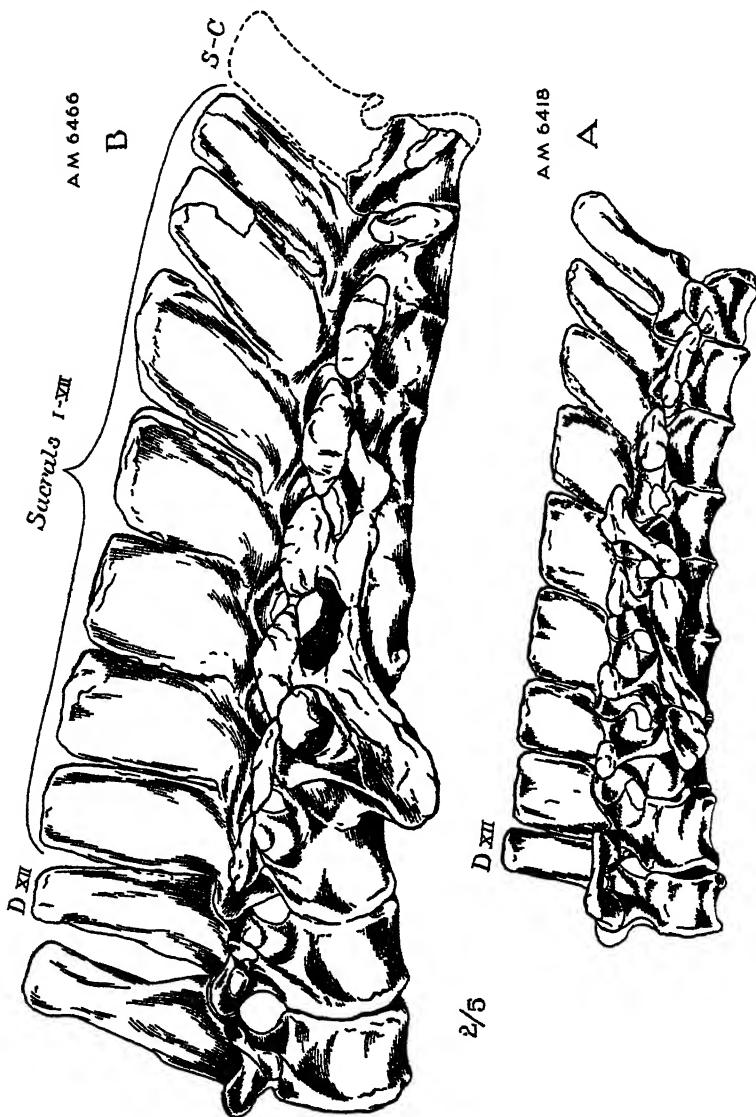


FIGURE 26 Sacra of *Protoceratops andrewsi*. A, fairly young adult individual. B, fully adult individual.

neural arch and partially on the centrum, about midway back on the vertebra. The whole surface of the end of this projection bears a sacral rib, approximately twice its length, with which it is completely coossified. This same condition is true for the seventh sacral and the sacro-caudal, except that the projections and ribs of these vertebrae are somewhat shorter. The distal ends of these ribs are enlarged,

especially antero-posteriorly, and in the young adult are in contact with one another, thus forming, together with the united distal ends of the diapophyses of sacrals one to five, a bar-like structure. In the fully adult, however, their contact with one another is lost, and their distal ends become proportionately larger, but more rounded in form.

The centra decrease in size from the first to the last sacral. Ventrally the first four are broadly crest-like, and the last three are definitely rounded. The centrum of the sacro-caudal, or eighth sacral, is somewhat crest-like ventrally, and is broader than the seventh.

In all of its important features, the sacrum of *Protoceratops* is definitely more primitive than that of any other known ceratopsian. Throughout its form and structure are to be seen precisely those characters to be expected in an archaic bipedal form which has just become quadrupedal. From it the sacrum of all the more progressive forms could have been derived, and the main changes that have taken place in the sacrum of these later members result primarily from a more complete adaptation to the quadrupedal habit, correlated with which is the flattening and lateral expansion of the ilia. The more important of these changes are best shown in a late Cretaceous genus such as *Triceratops*. They are as follows:

1. Two additional caudals have been taken over into the sacrum, thus bringing the total number of vertebrae incorporated in the sacrum of a fully adult to ten. At least, this is unquestionably the case in *Monoclonius*. There is, however, some question as to whether or not one of the two additional sacrals in *Triceratops* has been taken over from the dorsal series. According to the figures by Hatcher, Marsh, and Lull (1907, figures 53-55) of a *Triceratops prorsus* sacrum in the National Museum, the total number of coössified centra is ten. But a discrepancy lies in the fact that if the first of these corresponds to what we designate as the first sacral in *Protoceratops*, and to what has been designated as the first sacral in *Monoclonius* and the other known ceratopsians, then the first sacral parapophysial rib arises at the posterior end of the second sacral instead of at the posterior end of the first, as in all other recorded ceratopsians of which the details of the sacrum are known. Without any implication that "migration" of the acetabular bar cannot take place, we are of the opinion that this so-called first sacral in *Triceratops* is in reality the twenty-first presacral the centrum of which has become coössified with that of the first sacral. Our reasons for this conclusion are the following. In the first place, as pointed out above, the first parapophysial sacral rib arises at the posterior margin of the twenty-second vertebra in the

column, or the first sacral, in all of the other ceratopsians, and the number of presacrals is always twenty-one.* Secondly, the anterior portion of the *Triceratops* ilium is straighter and broader than in any other known form. It is, therefore, quite probable that one of the dorsals would become coössified with the sacrum. As shown above, there is a suggestion of this in the fully adult *Protoceratops*, which certainly seems correlated with the relatively longer pre-acetabular portion of the ilium in an individual of that growth stage. If, then, the so-called first sacral of *Triceratops* is in reality the last true dorsal, that would mean that only one additional caudal has been incorporated in the sacrum, and that there are only nine sacrals—or nine vertebrae which are definitely called sacrals in the ceratopsians. This is of no particular importance since the number of caudals incorporated in the sacrum of other later ceratopsians is somewhat variable. In *Styracosaurus*, for example, the tenth sacral is incompletely coössified with the ninth, and in *Pentaceratops*, what is definitely known to be the twenty-first of the presacral series is completely coössified with the first sacral and a tenth sacral is formed by the taking over of another caudal. This brings the total number of vertebrae functioning as a sacrum to eleven in this genus. What is important, is that in every known case the first parapophysial sacral rib appears at the posterior margin of the twenty-second vertebra in the column. There is no real reason for believing that the condition is different in *Triceratops*. Taking into consideration, therefore, all of the available information, the following conclusions, concerning the formation of the adult ceratopsian sacrum, seem probable. (a) The first parapophysial rib is always located at the posterior margin of the twenty-second vertebra and is mostly on the centrum of the twenty-third. (b) This rib is the first of four (possibly sometimes five as is suggested in *Pentaceratops*) which form the acetabular bar, and which are always located on vertebrae 23–26 inclusive. (c) These vertebrae probably are the only "true" sacrals in the Ceratopsia, and, therefore, represent the sacrum of the unknown, early Mesozoic ornithopodian ancestor of the suborder. (d) The twenty-second vertebra is always coössified with the twenty-third,—a condition that also may have been present in the unknown ancestral form. (e) The centrum of the twenty-first (or last "dorsal") is occasionally partially (*Protoceratops*) or wholly

* The number of free presacral vertebrae in *Triceratops* is not definitely known. Hatcher, Marsh, and Lull figure twenty-one (1907: 47), but the entire centra of the last five are missing and they were not joined to a sacrum. The sacrum described and figured by them is from another specimen. It is entirely possible that the neural spine of the last presacral of the figured column had a centrum that was coössified with the first true sacral. The number of presacrals in this genus would, therefore, be twenty-one.

(*Triceratops*, *Pentaceratops*) coossified with the twenty-second. (f) The minimum number of caudals added to the "true" sacrals is three, and the maximum, and usual number, in the later forms is five. In *Styracosaurus*, however, the fifth is incompletely coossified with the fourth, and in *Triceratops* only four are added.

2. When seen from above, the *Triceratops* sacrum is oval in outline, and the greatest diameter is across the parapophysial ribs of the fourth "true" sacral vertebra (the last of the four comprising the acetabular bar). In *Protoceratops* the sides of the sacrum are straight and the greatest width is across the parapophysial ribs of the first "true" sacral (the first of the four comprising the acetabular bar), behind which it becomes progressively narrower.

3. The whole sacrum is relatively short, broad, and upwardly arched, whereas in *Protoceratops*, it is long and narrow, and is rather straight when seen in lateral view.

4. In *Triceratops* the form of the acetabular bar is a high arch, and the distal ends of the ribs, of which it is composed, are completely coossified. In *Protoceratops* the acetabular bar has the form of a low arch, and coossification of the ribs is not as complete.

5. There is a pronounced median ventral groove on the centra that begins on the first "true" sacral and extends to the first free caudal (vertebrae 23 to 30 inclusive). In *Protoceratops* there is not even a suggestion of such a groove.

6. The neural spines are low and completely coossified. In *Protoceratops* they are high and, although closely applied, are not coossified.

In its general structure, the sacrum of *Leptoceratops*, while still very close to that of *Protoceratops*, shows a tendency towards all of these changes. The sacrum of *Brachyceratops*, so far as it is known, has progressed more in the direction of the more progressive forms. This is particularly true in the number of coossified vertebrae, which was undoubtedly at least eight, and probably nine, in the adult; in the low spines; and, in the absence of ventral, median keels on the centra. The sacrum of *Monoclonius* represents a further advancement towards the sacra of the most specialized forms of the latest Cretaceous, and is quite ideally intermediate between that of *Brachyceratops* and that of *Triceratops*.

Caudals

The caudal series of vertebrae is not completely known in any specimen. The greatest number articulated and completely preserved is thirty-two,* in specimen Am. Mus. No. 6417, a mounted skeleton. The total number was probably well over forty.

* Considering the number of sacrals as eight



P. andrewsi Group of two skeletons and a nestling. Assembled by Charles J. Lang. Skeleton in the foreground (Am Mus. No. 4417) mounted by V. M. Ling. Skeleton in the background (Am Mus. No. 6167) mounted by Peter C. Kuschn

BROWN AND SCHLAIKER PROTOCERATOPS



Protoceratops andrewsi. Two views of the same group shown in plate ten

As in *Leptoceratops*, the neural spines are very tall, and are erect in relation to the long axis of the centra. The spines on the first four caudals are expanded antero-posteriorly almost as much as those of the posterior sacrals. On five and six they are less expanded, and on the remaining vertebrae the spines become more oval-shaped to rounded in cross-section. From the first to the fourteenth they become progressively taller, and then become gradually shorter posteriorly. The very tall neural spines of the caudal vertebrae show that the supracaudal muscles were deep and powerful. This feature of the tail is also dominant in *Leptoceratops*,—possibly an adaptation for a swimming habit.

The transverse processes, which are composed of ribs more or less completely ossified with the centra, are present on the first thirteen caudals. They are rather short on the first caudal, somewhat longer on the second, and longest on the third. They then become gradually shorter on the succeeding vertebrae, and are but mere welts on the thirteenth caudal. They are always directed outward and slightly backward, are always attached at the sides of the centrum where it unites with the neural arch, and at their bases are expanded antero-posteriorly the length of the arch.

In the anterior caudals, the width, length, and depth of each centrum are subequal. From the twenty-fifth on, however, the centra are definitely longer than they are wide or deep. The first chevron is present at the union of the third and fourth caudals. Chevrons three to nine are the longest and heaviest, and from the tenth on they become gradually smaller.

When the caudal vertebrae of *Protoceratops* are compared with those of the other ceratopsians, the closest similarity, and indeed a striking one, is seen between this genus and *Leptoceratops*. In the caudal vertebrae of the more progressive forms, several distinct changes have taken place. The neural spines tend to become directed posteriorly with respect to the centra, and become reduced in height. Transverse processes are present on a greater number of vertebrae and are located farther down on the centra. The centra become deeper and wider than they are long. In these features *Triceratops* represents the extreme. *Monoclonius*, however, seems less specialized in these characters. The neural spines are still quite erect, and while they are rather short, they do not become progressively shorter after the first few caudals. The first four spines are the tallest, five and six are a little lower, and seven to fifteen are more or less uniform in height,—at least as is shown in the very complete and unusually well

preserved skeleton of *Monoclonius nasicornis* in the American Museum (No. 5351).

COMPARATIVE STUDY OF THE RIBS AND STERNUM

Ribs

Ribs are present on all vertebrae except the posterior caudals. The first five cervical ribs are short, and of these, the first two are the longest. The first, or atlas rib, has the tubercle only slightly developed, and on the ninth cervical rib the capitulum and tubercle are more widely separated than on any rib. The third to the sixth dorsal ribs are the longest. The third and fourth are subequal in length, and the fifth and sixth are somewhat longer. Of the whole dorsal series of ribs, the eleventh or last of the free ribs, is the shortest, and the twelfth is very much reduced in length and is in contact with the antero-inferior surface of the ilium. In the fully adult individual, however, as shown in Am. Mus. No. 6466, since the front of the ilium extends proportionately even farther forward, the eleventh becomes shortened and extends down and under the front of the ilium. The character of the distal ends of the anterior dorsal ribs, and indentures on the posterior margins of the sternal plates, strongly suggest that cartilaginous abdominal ribs were present.

The sacral and caudal ribs have been described under the consideration of the sacrum and the caudal vertebrae.

There are no striking differences between the ribs of *Protoceratops* and those of the later ceratopsians. In the later forms the dorsal ribs become somewhat proportionately longer and become more curved. The latter feature is correlated with the development of more upwardly directed transverse processes of the dorsal vertebrae.

Sternum

Sternal plates are preserved in three skeletons of *Protoceratops* (Am. Mus. Nos. 6416, 6417, and 6467). They are relatively long and narrow, and in transverse section are slightly convex ventrally. Each has a slightly concave outer margin and a slightly convex inner margin. Anteriorly they are broadly rounded, and posteriorly they are rather narrow and gradually flare outward. Slight indentures are present on the somewhat thickened posterior margins, which suggests that cartilaginous abdominal ribs were present.

Compared with the sternal plates of *Monoclonius* and *Triceratops*, those of *Protoceratops* are quite primitive. In the later forms, the

following changes take place. They become proportionately broader; the external borders become more deeply concave; the antero-external margins become more pointed; and, posteriorly they are relatively broader, and flare more abruptly outward.

COMPARATIVE STUDY OF THE APPENDICULAR SKELETON

Pectoral Girdle

The pectoral girdle is completely known in a dozen or more specimens representing several distinct growth stages from the very immature to the fully adult individual. It consists of three paired elements,—scapula, coracoid, and clavicle. The clavicle has not been recorded previously in the Ceratopsia, but as will be shown later, there is reason for believing that it probably was present in all of the described forms in which a scapula and a coracoid are known.

SCAPULA

In its general form, the scapula is long and slender. The mid-section, including about one-third of the entire length, is narrow, and oval in cross-section. The upper end is flattened, thin, and well expanded antero-posteriorly. The broadest and thickest portion of the scapula is just above the glenoid cavity. The antero-ventral margin is thin and is extended downward. Thus, the ventral margin faces downward and obliquely backward with respect to the long axis of the bone. Less than one-half of the glenoid cavity is formed by the scapula. A low ridge extends from the upper margin of this cavity obliquely across the outer surface of the scapula and becomes confluent with the anterior margin about one-third of the way down from the upper end. This ridge separates the posterior area that gives origin mainly to the *deltoides scapularis* muscle, from the anterior area which gives origin mainly to the *scapulo-humeralis* muscle. It cannot be homologous with the spine of the scapula in mammals, as suggested by Hatcher, Marsh, and Lull (1907: 58). As shown by several who have dealt with the musculature of the shoulder girdle, especially Romer (1922), the *infraspinatus* and *supraspinatus* muscles, which occupy most of the areas on either side of the spine of the mammalian scapula, are homologous with the *supracoracoideus* muscle in the Reptilia, and the spine of the mammalian scapula is homologous with the ridge along the anterior margin of the scapula in the mammal-like reptiles.

From the very immature to the fully adult stage, several marked changes take place in the scapula. (See FIGURE 26.) It becomes relatively broader and thicker, especially in the portion just above the glenoid cavity. The upper border of the glenoid cavity becomes relatively heavier, and more extended posteriorly. The upper end of the scapula curves forward, and the whole anterior margin becomes rather deeply concave. The ridge on the outer surface occupies a more central position. When viewed from the front, the scapula of the immature individual is quite straight, but in the older individual, it is pronouncedly curved being outwardly convex.

All of these changes foreshadow the main modifications which take place in the scapula of the later ceratopsians. In these later forms two other changes take place. The ventral margin of the scapula becomes horizontal with respect to the long axis (*Triceratops*), and more than half of the glenoid cavity is formed by the scapula, whereas in *Protoceratops* over half of the glenoid cavity is formed by the coracoid. In these features, and in those given above, the scapula of *Monoclonius* is more primitive than that of *Triceratops*.

CORACOID

In its general form, the coracoid of *Protoceratops* is very similar to that of the other ceratopsians. It differs in some details, all of which seem to be primitive for the group. Compared with the coracoid of *Triceratops*, it displays the following main differences. It is relatively deeper, and the ventral border is more curved. The ventral posterior projection is more pointed, and it is neither as broad nor as thick. The posterior notch between this projection and the glenoid cavity is deeper, and there is a large pit located at the outer margin in the deepest portion of this notch—a character which may signify that the costo-coracoideus muscle was especially well developed. Slightly more than half of the glenoid cavity is formed by the coracoid, and the whole cavity is more outwardly directed. This latter feature signifies that in its normal position the humerus was more anteriorly directed than in the more advanced forms. In all of these features the coracoid of *Protoceratops* is extremely close to that of *Leptoceratops*. That of *Monoclonius* is much more advanced, but is considerably more primitive than that of *Triceratops*. (See FIGURE 26.)

CLAVICLE

Clavicles are completely preserved in several specimens of *Protoceratops*. In each of these specimens, this element was found lying

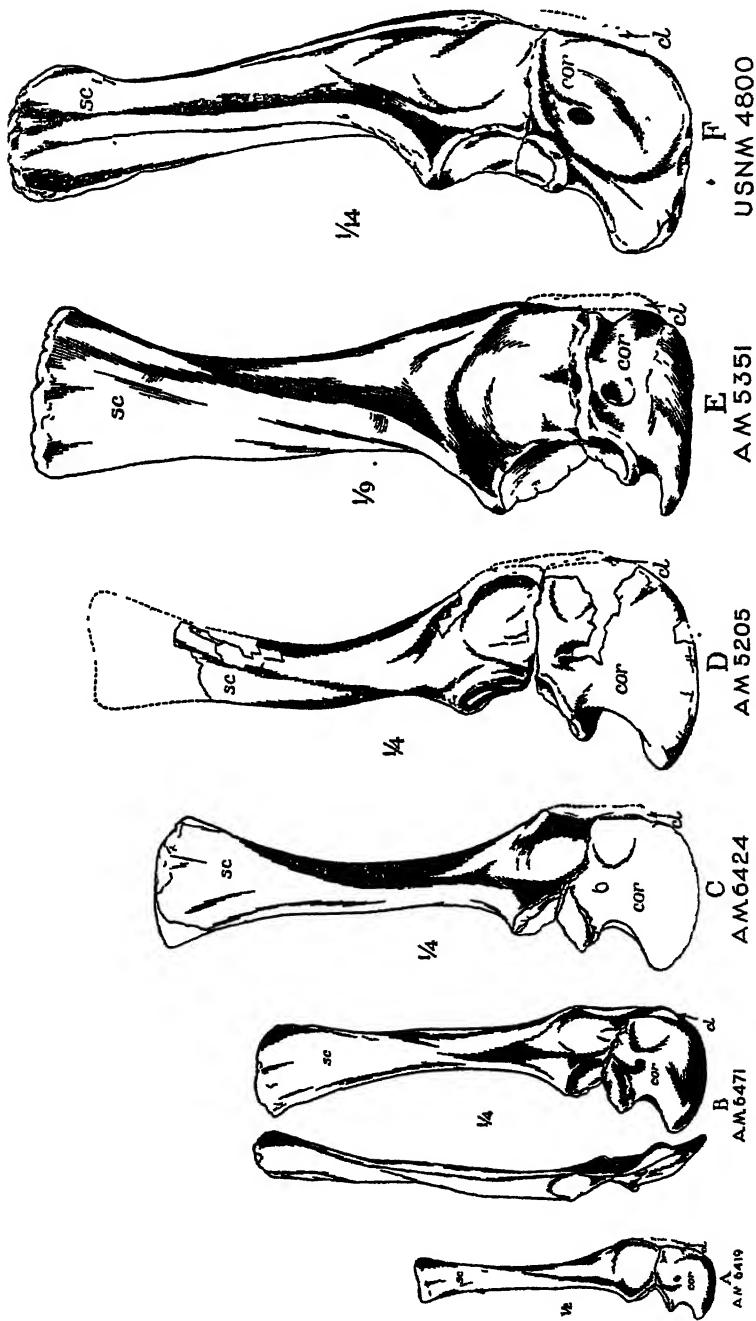


FIGURE 26. Comparative series of ceratopid pectoral girdles. A-O, *Protoceratops andrewsi*, representing three growth stages from $\frac{1}{4}$ very immature to a fully adult individual. D, *Lepoceratops gracilis*. E, *Monoclonius (Centrosaurus) nasutus*. F, *Triceratops prorsus*, modified from Hatcher, Marsh, and Lull.

against the ventro-anterior margin of the scapula and the dorso-anterior margin of the coracoid. This seems to be its natural position, for it articulates perfectly with these margins.

In its general form, the clavicle is short and rod-like. Its ventral portion is turned inward so that when seen from the front its outer margin is markedly convex, and its inner margin concave. The upper end is pointed, and is thicker behind than in front. The lower end is broad and rugose, and is thinner behind than in front. This peculiar shape results mainly from an inward twisting of the ventral portion, thus causing what is the anterior margin in the upper portion to become the anterointernal margin in the lower portion.

A clavicle has not been reported previously in the Ceratopsia. However, there is reason for believing it was present in all of the recorded forms. In *Protoceratops* the ventro-anterior margin of the scapula, and the dorso-anterior margin of the coracoid are straight and rugose where they unite with the clavicle. These features are characteristic of the scapulae and coracoids of all the known forms in which these elements have been recorded. It is probable, therefore, that clavicles were present in all of the ceratopsians, and for this reason we have restored them on the shoulder girdles of the later forms illustrated in FIGURE 26.

Fore Limb

The most outstanding feature of the fore limb is its relatively small size. It is scarcely more than one-half the length of the hind limb. In all ceratopsians the fore limb is shorter than the hind limb, but in no other genus is the discrepancy as great as in *Protoceratops*. Even in *Leptoceratops*, which in most of its features is closer to *Protoceratops* than to any other ceratopsians, the fore limb is almost three-fourths the length of the hind limb. In an end-member of the group, such as *Triceratops*, the difference is even less, and in *Monoclonius*, the relative difference in length of the fore and hind limbs is intermediate between *Brachyceratops* and *Triceratops*. The short fore limbs of *Protoceratops* together with other characters displayed in the pelvic girdle and hind limbs, show that this primitive ceratopsian was still close to the earlier, bipedal ornithischian ancestor of the group. The proportionate lengthening of the fore limbs in the later ceratopsians is correlated with a more complete adaptation to a quadrupedal habit.

HUMERUS

The humerus is relatively long and slender. This is particularly true in the very immature individual. The proximal extremity is

fairly broad, and is deflected inwardly—more so in the adult than in the immature specimens. The proximal articular surface becomes more distinct with age. The head is located at approximately the middle of this surface. It is crest-like in the young, but becomes rounded in the adult. The delto-pectoral crest is long and thin in the immature stage, and is marked off from the articular surface by a fairly deep notch. With age, it develops into a thick, blunt projection quite far removed from the proximal end of the humerus. In the early growth stage the shaft is triangular in cross-section—being broad behind and narrow in front. In the adult stage, it is definitely rounded in cross-section, which is due mainly to the reduction of the ridge that extends from the delto-pectoral crest down the front of the shaft. The distal extremity is relatively narrow. The size of the condyles is limited, and as in all ceratopsians, the entepicondyle is situated slightly below the ectepicondyle.

The humerus of *Protoceratops* is distinctly primitive, as is shown by its elongated and slender form, by the narrowness of its extremities, and by the proximal position of the delto-pectoral crest. In the later ceratopsians the humerus undergoes considerable modification. It becomes a relatively short and massive element, with its extremities widened and heavy, its shaft short, and its strong delto-pectoral crest extending downward two-thirds its entire length. These modifications are correlated with the more thorough adaptation to a quadrupedal habit, which necessitates the development of a type of humerus adapted for supporting greater weight. A tendency towards these modifications is also expressed in the development of the humerus of *Protoceratops* from the immature to the fully adult stage. (See FIGURE 27).

ULNA

The ulna is typically ceratopsian in its general form. Compared, however, with that of a later form, such as *Triceratops*, it is considerably more slender, and its extremities are not as expanded. Also, the olecranon is relatively not as high, and the proximal articular surface faces more upward, rather than forward. In these characters, the ulna of *Protoceratops* is definitely more primitive. Its closest affinities are to be seen in *Leptoceratops* and *Brachyceratops*—particularly in the former. The ulna of the very immature *Protoceratops* is even less like that of the later forms than is the ulna of the adult. In *Brachyceratops*, the only ulna known is from a young individual. Perhaps this accounts for its close similarity to that of *Protoceratops*. In the adult, it probably was more like the ulna of *Monoclonius*.

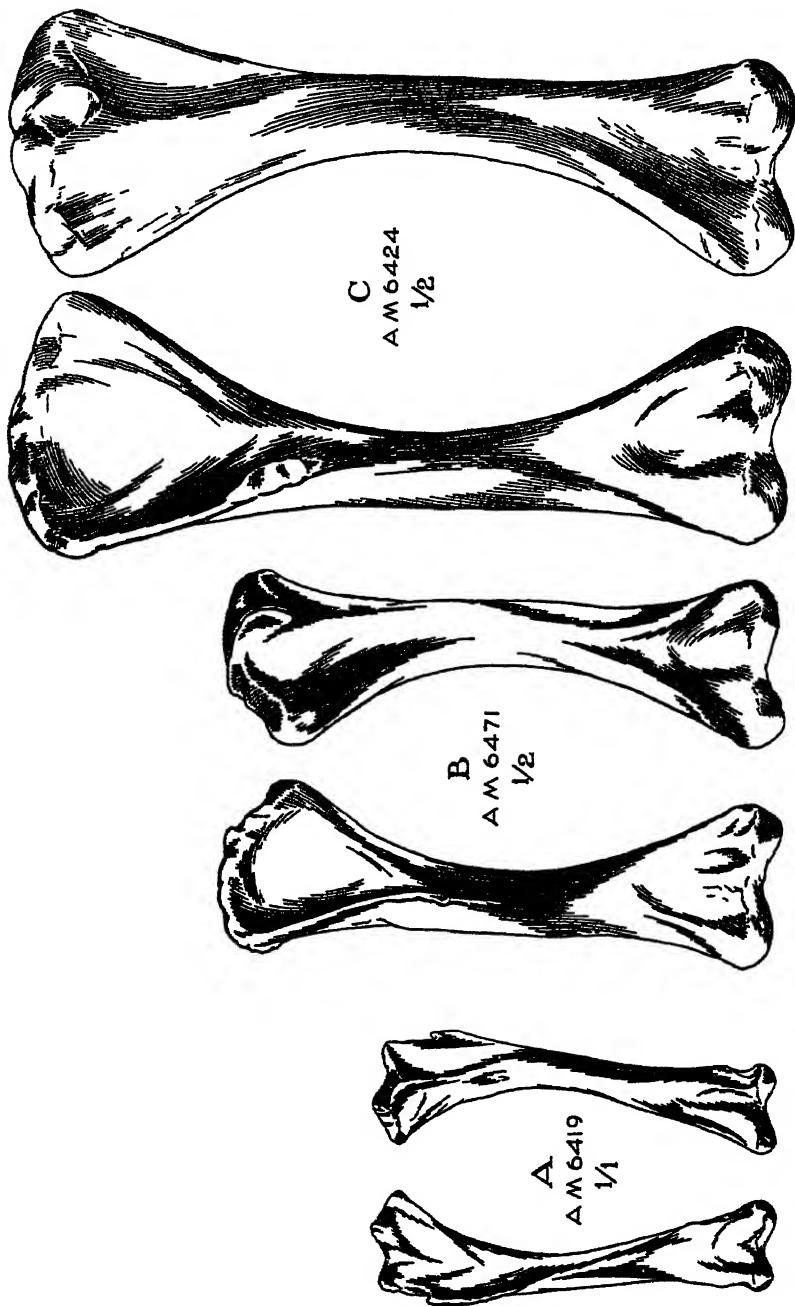


FIGURE 27. *Protoceratops andrewsi*. Anterior and posterior views of humeri representing three growth stages from a very immature (A) to a fully adult (C) individual

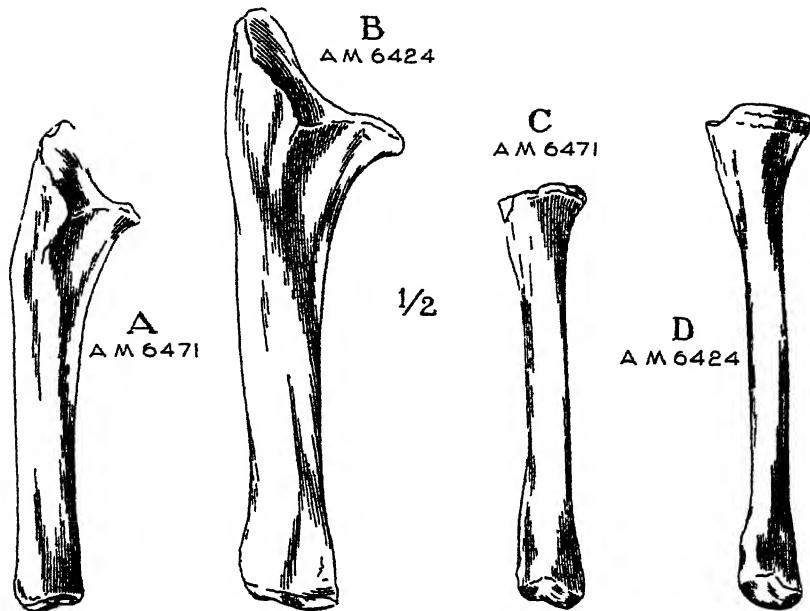


FIGURE 28. Ulnae and radius of two *Protoceratops andrewsi* specimens. A and C, young adult; B and D, fully adult.

RADIUS

In keeping with the archaic features of the ulna, the radius is also proportionately more slender, and its extremities are relatively less expanded than in any of the other ceratopsians. With age, however, it is less distinctive in these features, and resembles more closely the radius of the more progressive forms. The radii of a young adult and a fully adult individual are shown in FIGURE 28.

MANUS

In all known ceratopsians, the manus is smaller than the pes, but in *Protoceratops* the discrepancy is greater than in any of the other forms. For example, the third metacarpal is less than one-third the length of metatarsal III, whereas in *Styracosaurus* metacarpal III is two-thirds the length of metatarsal III.

The carpus of *Protoceratops* compares favorably with that of *Leptoceratops*. The form and proportions of the ulnare and radiale are approximately the same as in the latter genus. At the proximal ends of metacarpals III and IV, and wedged between the ulnare and radiale, is a rather large, somewhat keystone-shaped element which probably

is the intermedium. That it may represent the second carpal is possible, however. This element is also present in *Leptoceratops*, although it is quite reduced in that genus. It has not been recorded in any of the later ceratopsians. Two carpals, probably the third and fourth, are preserved and are located above metacarpals III and IV, as in all of the other ceratopsians in which these elements are known.

As in all ceratopsians, metacarpals II and III are subequal in size, and metacarpal I is about two-thirds the length of II or III. In the later forms, metacarpal IV is approximately the same size as I, and V is only slightly smaller than I or IV. In *Protoceratops*, however, IV is only about three-fourths the length of I, and V is only one-half the size of I.

The phalangeal formula is 2, 3, 4, 3, ?2. Only one phalanx is preserved on the fifth digit, but its distal end shows a well formed articular surface, which probably signifies that a second was present as in *Styracosaurus* and *Monoclonius* (*Centrosaurus*). Only one is preserved in *Leptoceratops* but its distal end also shows a distinct articular surface for a second.

The phalanges are typically ceratopsian. They are, however, more slender than in the progressive forms, in which character they are closer to those of *Leptoceratops*. Ungual phalanges are present only on the first three digits,—a constant feature for the whole group. The first is the largest and the third is the smallest. The same holds true for *Leptoceratops*, but in the later forms they are subequal in size. In general form, they are quite elongated and pointed, but they definitely show a tendency towards broadening and shortening, which is so characteristic in the large and more progressive forms. In *Leptoceratops*, however, they are longer, more pointed, and much more arched—being, therefore, in these respects, more primitive than those of *Protoceratops*. The dorsal surface of each is quite convex longitudinally and transversely, and each postero-lateral margin is pierced by a foramen. In the later forms these foramina at first become elongated grooves on the margins (*Leptoceratops*), and later open notches (*Monoclonius* (*Centrosaurus*), *Styracosaurus*).

Considering the manus of *Protoceratops* in its entirety, it does not possess a single character that is not either typically ceratopsian or potentially so. It is definitely primitive and is, therefore, extremely like that of *Psittacosaurus* in many of its characters.

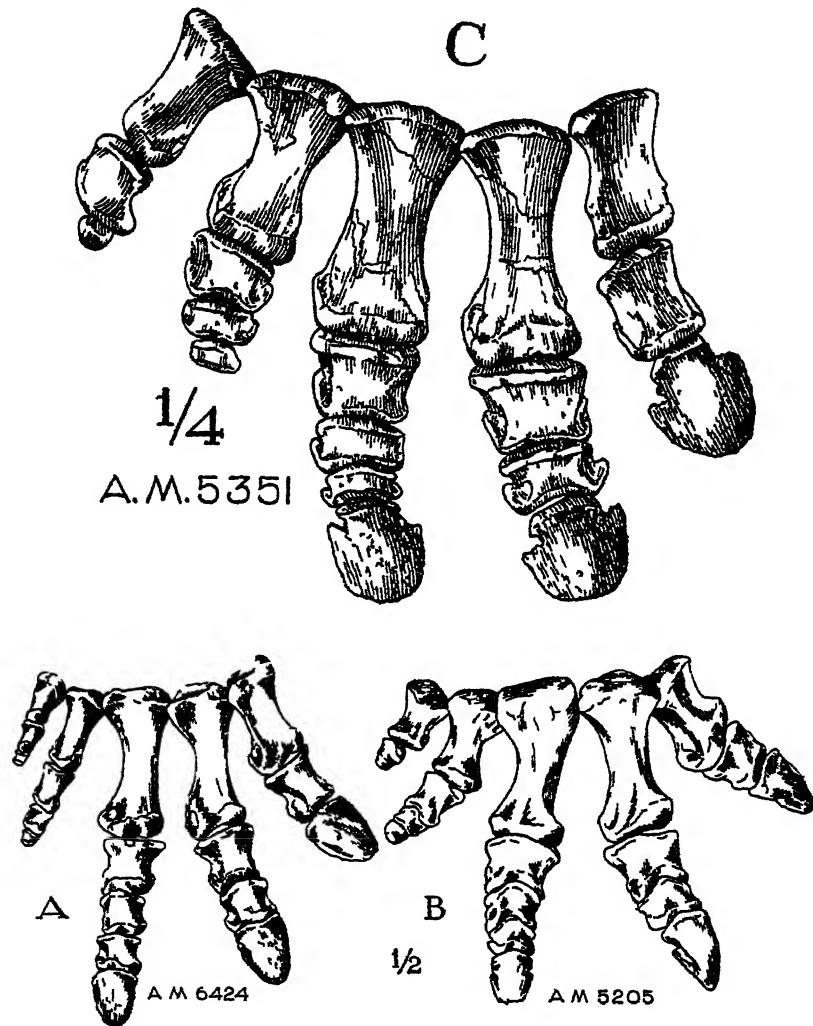


FIGURE 29 Comparative illustrations of the right manus of three ceratopsians A. *Protoceratops andrewsi* (fully adult) B. *Leptoceratops gracilis* C. *Monoclonius (Centrosaurus) nasicornus*

Pelvic Girdle

The pelvic girdle of *Protoceratops* is more primitive than that of any other ceratopsian. This is especially shown by such outstanding features as the erect position of the dorsal margin of the ilium; the very small prepubis; and, the straight, long, and slender ischium. In the more progressive forms, the dorsal margin of the ilium becomes

everted; the prepubis becomes elongated, and the distal end much expanded dorso-ventrally; and, the ischium becomes heavy, downwardly curved, and shortened. These, and many other less salient characters of the *Protoceratops* pelvic girdle—to be considered in detail below—show that this archaic ceratopsian was still essentially a bipedal form in its pelvic structure. The changes that take place in the pelvis of the later forms are obviously correlated with a more complete adaptation to a quadrupedal habit. Evidence of this fact is shown by Romer in his illuminating paper on “The Pelvic Musculature of Ornithischian Dinosaurs” (1927).

ILIUM

The most outstanding single feature of the ilium is the erect position of the dorsal margin. It shows, however, in the anterior projection, the beginning of the eversion of this margin, which is so characteristic of the ilium in the later forms. The whole of this projection curves outward, and the front of its dorsal margin is turned slightly outward. There is quite a broad ventral shelf for the pubo-ischio-femoralis internus, and the anterior end is rugose for the attachment of axial muscles. Extending over most of the lateral surface, and including the dorsal margin, of this anterior projection, is a very clearly marked area which undoubtedly served as the origin for a distinct ilio-tibialis 1 (sartorius) muscle. Posterior to this the dorsal margin presents a well demarcated rugose area for the ilio-tibialis 2, and the remainder of the lateral surface shows four distinct zones of attachment. The anterior, and main portion of this surface shows two distinct areas, of which the first undoubtedly gave origin to the ilio-trochantericus, and the second to the ilio-femoralis. Below and behind the latter is a fairly large area for the ilio-fibularis, and on the postero-ventral margin there is a distinct rugose area for the flexor tibialis externus muscle. On the posterior margin there is a distinct zone of attachment for part of the dorsal caudal musculature. The postacetabular portion of the ilium has only a slight broadening of the ventral margin anteriorly. Perhaps part of the coccygeo-femoralis brevis originated from this area.

Ilia representing early growth stages are not known. From the known young adult to the adult stage, however, certain marked changes take place. There is a proportionate increase in length of the region anterior to the postacetabular portion, and the anterior projection becomes more outwardly deflected, its ventral shelf broadens, and the anterior portion of its dorsal margin shows a stronger tendency to

become outwardly turned. All of these are features which find greater emphasis in the ilia of the later ceratopsians. In addition, the ischiac peduncle become relatively larger with age, and the depth of the ilium at this peduncle becomes proportionately greater. (See FIGURE 30.)

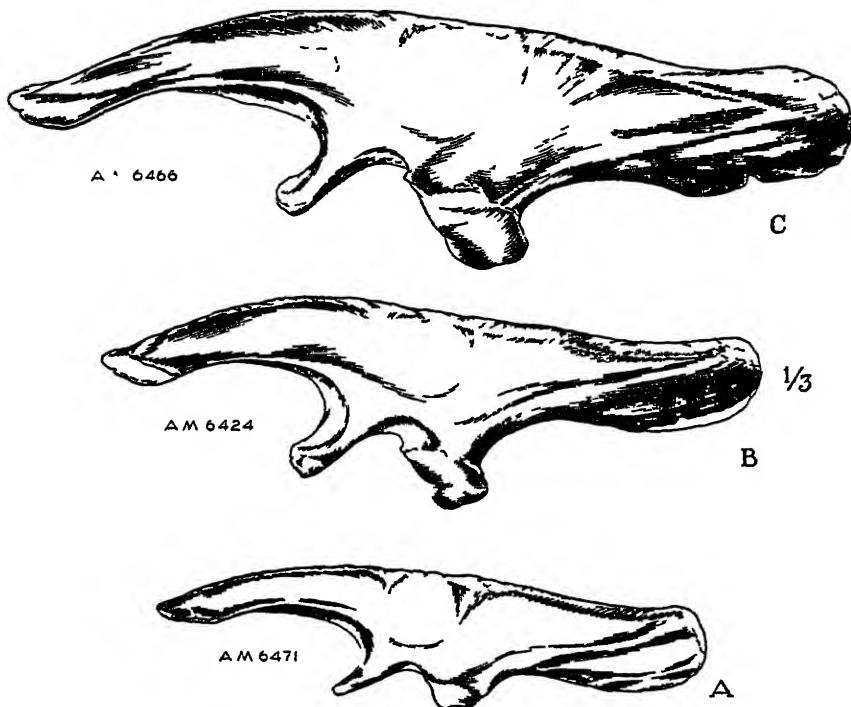


FIGURE 30. *Protoceratops andrewsi*. Ilium representing three growth stages from the young adult (A) to the fully adult (C) individual.

To derive the ilia of the later ceratopsians from that of *Protoceratops* would require only a continued outward turning of the dorsal margin, and a continued broadening of the anterior ventral shelf and of the postacetabular ventral margin. Of the later forms, the ilium of *Leptoceratops* is closest to that of *Protoceratops*. In all of its important characters, however, it is definitely more progressive, but is still closer to the ilium of *Protoceratops* than to that of any of the other known forms.

PUBIS

The most significant feature of the pubis is the short anterior process, or prepubis. (See FIGURE 31.) With age, however, there is a tendency

for it to become proportionately longer, and for the distal end to become dorso-ventrally expanded. The anterior margin is rugose, which shows that the main function of the prepubis—as in all the ceratopsians—was probably abdominal support. Its shortness is a heritage character—a dominant feature of the early ornithischian prepubis—and probably was correlated, in part at least, with the short abdomen.

The acetabular portion is deep and heavy, and gives off the posteriorly deflected portion (= true pubis) at its anterior ventro-inner margin. This posterior projection originates farther forward on the acetabular portion, curves inward farther at its base, and is proportionately longer than in any of the other ceratopsians. The obturator foramen is large. At its postero-ventral margin there is a shallow notch, against which fits the antero-ventral margin of the pubic process of the ischium.

The pubis of *Protoceratops* is more primitive than that of any other ceratopsian. This is principally shown in the following characters:

1. An extremely short anterior process, or prepubis.
2. The anterior position at which the posterior process originates, the marked inward curvature of its basal portion, and its unusual length.
3. The large size of the obturator foramen.

While the pubis of *Leptoceratops* is very similar to that of *Protoceratops*, it is somewhat more advanced in all of its features.

ISCHIUM

The ischium of *Protoceratops* is more primitive than in any of the other known ceratopsians. Its closest affinities are with the ischium of *Leptoceratops*, but in this form, that element, especially in its robustness and in its curvature, is closer to that of *Brachyceratops*.

The most apparent unique features of the *Protoceratops* ischium are its relatively straight form, its slenderness, its relatively great length, and its somewhat expanded distal end. (See FIGURE 31.) In the more progressive forms, the ischium becomes very much downwardly curved, proportionately robust and shortened, and its distal end assumes more slender and pointed proportions. The relationship of these changes to function has been discussed by Romer (1927: 252-253).

The still somewhat expanded distal end of the *Protoceratops* ischium suggests a strong ischio-caudalis. This corroborates the evidence shown by the tall neural spines on the caudal vertebrae that the tail possibly was used for swimming. On the internal surface of the dorsal

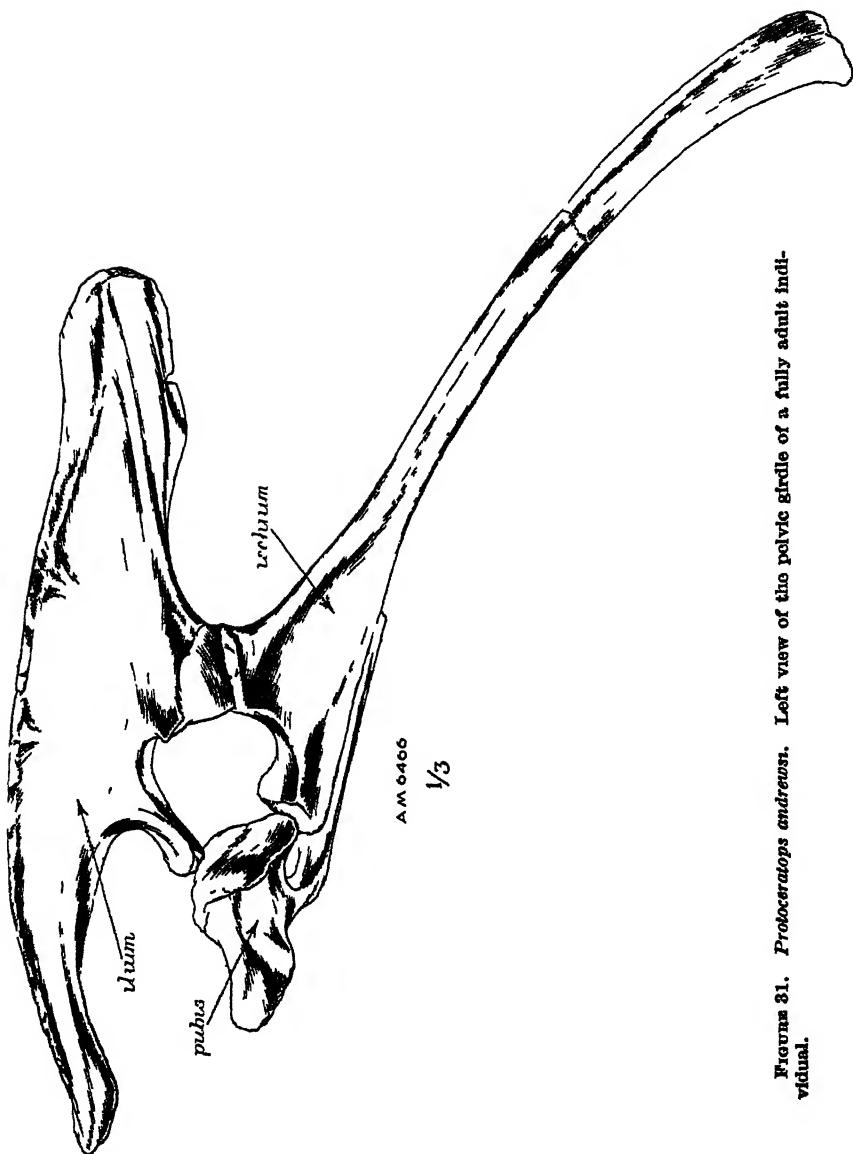


FIGURE 31. *Protoceratops andrewsi*. Left view of the pelvic girdle of a fully adult individual.

margin there is a broad shallow groove which extends from near the base of the enlarged anterior end back more than halfway to the distal extremity. This probably marks the origin of the ischio-trochantericus.

Hind Limb

While every element of the hind limb is definitely more primitive than in any of the later ceratopsians, perhaps the most striking single feature is that the proximal elements are proportionately short and the distal elements are proportionately long. In TABLE 3 measure-

TABLE 3
COMPARATIVE LIMB RATIOS OF *Psittacosaurus*
"*Protiguanodon*," AND FIVE CERATOPSIAN GENERA

Genus	Length of femur mm.	Length of tibia mm.	Length of mts. III mm.	Femoro-tibial ratio	Femoro-mts. III ratio	Tibio-mts. III ratio
<i>Psittacosaurus</i>	162	179	?	1.104	?	?
Am. Mus. 6454						
" <i>Protiguanodon</i> "	155	158	78	1.019	.503	.493
Am. Mus. 6253						
<i>Protoceratops</i> ♂ (young adult)	189	208	96	1.100	.508	.461
Am. Mus. 6471						
<i>Protoceratops</i> ♂ Am. Mus. 6417	221	241	115	1.090	.520	.477
<i>Protoceratops</i>	226	241	119	1.066	.526	.493
Am. Mus. 6416						
<i>Protoceratops</i> ♂ (adult)	248	273	125	1.100	.504	.457
Am. Mus. 6424						
<i>Leptoceratops</i>	346	355	?	1.023	?	?
Am. Mus. 5464						
<i>Brachyceratops</i>	337	268	97	.795	.287	.362
U.S.N.M. 7953						
<i>Monoclonius</i>	740	553	215	.747	.290	.388
Am. Mus. 5351						
<i>Triceratops</i>	980	580	253	.591	.258	.436
Am. Mus. 5083						

ments of the femur, tibia, and the third metatarsal, together with ratios of these elements, are given for *Psittacosaurus*, "*Protiguanodon*," four specimens of *Protoceratops andrewsi* representing individuals from the young adult to the adult growth stage, and four genera of later ceratopsians. *Protoceratops* and *Leptoceratops* are unique among the Ceratopsia in having the femur shorter than the tibia, and in that the foot is unusually large and long. The data presented in TABLE 3 shows clearly that in the later forms both the tibia and the foot become proportionately shortened. Furthermore, in all of them, except

Triceratops, the foot becomes shorter in relation to the tibia. In *Triceratops*, however, the tibio-metatarsal III ratio is close to that of the adult *Protoceratops*. The explanation of this is obvious. In the latter, the foot is exceptionally long, and in *Triceratops* the tibia is exceptionally short. Within the genus *Protoceratops* itself, a considerable amount of variation is displayed. This may be due, in part to

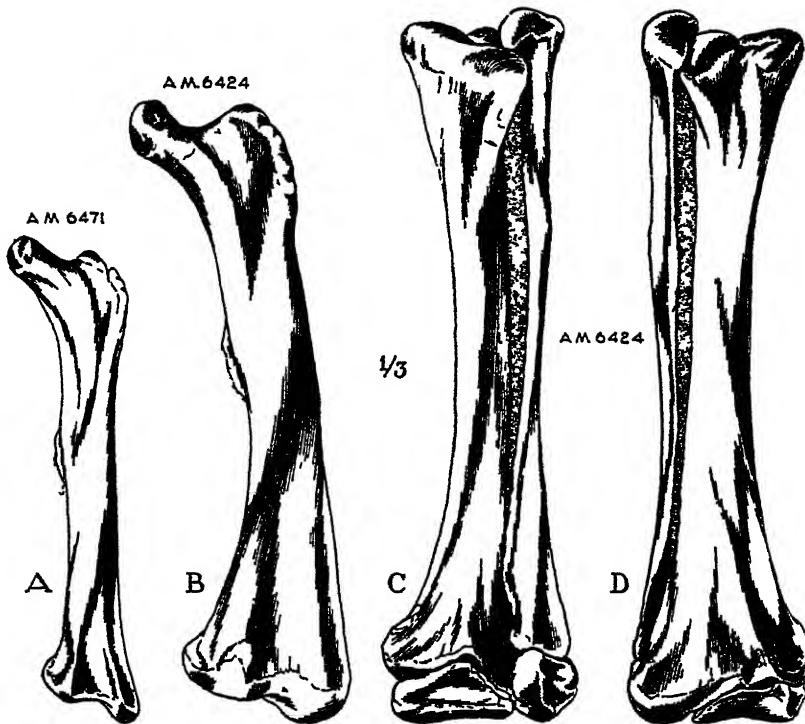


FIGURE 32. *Protoceratops andrewsi*. A, left femur of a young adult. B, left femur. C and D, anterior and posterior views of the left tibia and fibula, of a fully adult form.—fibula and calcaneum slightly displaced upward.

sexual variation, and in part to individual variation. It will be noticed, however, that three of the four specimens, for which the data are given, belong to the group which we regard as "males." It would seem, therefore, more probable that individual variation is responsible for these differences. The limb ratios give further emphasis of the primitiveness of *Protoceratops*. In all of them, this genus is closer to the two early Cretaceous bipedal ornithopods, *Psittacosaurus* and "*Protiguanodon*," and to *Leptoceratops* than to the later ceratopsians.

FEMUR

As shown above, the femur is shorter than the tibia, but is relatively slender when compared with the femora of later forms. The head is proportionately small, and faces inwardly more than in the progressive types. Its relationship to the acetabulum, therefore, indicates that the position of the limb was further under the body than in the later, more quadrupedal forms, in which the head of the femur faces more upward than inward, and in which there has been no appreciable change in the plane of the acetabulum.

The head is well differentiated from the shaft by a rather long neck, and the level of its uppermost margin is considerably above the "greater trochanter." The "lesser trochanter" is large and is distinctly marked off from the "greater trochanter" by a notch above, and by a shallow groove on the outer surface. The fourth trochanter is situated nearly halfway down on the shaft, and is large and pendant. The internal condyle of the distal extremity is considerably smaller than the external condyle, and it is located quite far above the ventral limit of the latter.

In the later ceratopsians the femur becomes rather modified in a number of characters, of which the following are the more salient:

1. The femur in general assumes a great robustness.
2. The head becomes proportionately larger, tends to become less differentiated from the shaft, and faces more upward than inward.
3. The "greater trochanter" and "lesser trochanter" tend to fuse, and to assume a position almost level with the upper margin of the head. This relative rise in position of these trochanters in the femora of the later ceratopsians seems correlated with the everted dorsal portion of the ilium, which position would place the ilio-trochantericus and the ilio-femoralis externus in a more vertical plane.
4. The fourth trochanter becomes reduced to a low protuberance.
5. The internal condyle of the distal extremity more nearly approaches the ventral limit of the external condyle.

In the *Brachyceratops-Monoclonius-Triceratops* line, these changes are less emphasized in *Brachyceratops* than in *Triceratops*. The femur of *Leptoceratops* is close to that of *Protoceratops*, but shows a definite tendency in the direction of the later forms. Unfortunately no femur in the *Protoceratops* collection represents an immature growth stage. It is interesting to see, however, that from the young adult to the adult stage several of the above listed changes in the later forms are foreshadowed. As shown in FIGURE 32, the adult stage displays a greater robustness, the "lesser trochanter" becomes more closely

affiliated with the "greater trochanter," and the lower limit of the internal condyle more nearly approximates that of the external condyle.

TIBIA

The tibia is long and slender, and is unique in having the proximal end but slightly expanded. (See FIGURE 32 C and D.) Apart from this feature, however, the proximal end is very similar to that of *Monoclonius*. The main body of the shaft is quite cylindrical in cross-section, and becomes flattened as it merges into the expanded distal end. As in all ceratopsians, its outer portion projects somewhat below the level of its inner portion so that in posterior view it is seen to extend down behind nearly all of the calcaneum. In *Triceratops* this outer portion of the distal end is unusually extended. According to Lull (1933: 60), in this genus, ". . . the calcaneum has either been replaced by, or absorbed into, . . ." this process of the tibia. The absence of a calcaneum in some specimens of *Triceratops*, in no way substantiates such a view. It simply means that this element probably was not preserved with the rest of the specimen. The right hind limb of the mounted *Triceratops* skeleton in the American Museum (No. 5033) shows a large calcaneum in an excellent state of preservation. It is situated in front of the enlarged, and quite extended external process of the distal end of the tibia, as is normal for the group. It was undoubtedly present in all ceratopsians.

As mentioned previously, the tibia of the later ceratopsians becomes proportionately much shortened. Along with this change, it becomes very robust, its proximal and distal extremities expand, and the shaft becomes flattened antero-posteriorly.

FIBULA

In conformity with the general features of the tibia, the fibula is long and slender. Apart from its slenderness, it is very similar to the fibula of the later ceratopsians, which has been thoroughly described in several genera. Its distal extremity, however, is less expanded, is more flattened, and is closely applied to the front of the tibia for a relatively greater distance. In these features, it is more like that of *Leptoceratops* than of any other genus. The fibula of *Leptoceratops*, however, seems equally as close to the somewhat more specialized fibula of *Brachyceratops*, as it is to that of *Protoceratops*.

PES

The tarsus is composed of four elements—a calcaneum, an astragalus, and two tarsalia in the distal row. Both the astragalus and cal-

caneum are similar to those of *Leptoceratops*, as described by Gilmore (1939: 7). They differ, however, in that the astragalus of *Protoceratops* is relatively larger and presents a pronounced median antero-dorsal projection. Also, the distal articular surface of the calcaneum of *Protoceratops* is considerably larger. In the more progressive ceratopsians, the astragalus becomes shallower, and, since the distal outer portion of the tibia becomes relatively enlarged, it seems to occupy less of the distal surface of the tibia. The calcaneum in these forms becomes closely adhered to, or fused with the astragalus, and becomes dorso-ventrally flattened.

The two distal tarsalia are best preserved in specimen Am. Mus. No. 6478. The inner one is half again as large as the outer one. Its upper articular surface is slightly concave antero-posteriorly and is as wide as the astragalus. Ventrally it articulates with the entire proximal surfaces of the first three metatarsals, and with the dorso-inner margin of metatarsal IV. On its anterior margin there is a projection which lies over the union of the second and third metatarsals. Its outer margin presents an extensive convex surface for articulation with the concave inner margin of the other tarsal. The dorsal surface of this second tarsal receives the large articular surface of the calcaneum. Ventrally this element articulates mainly with the concave surface of metatarsal IV, but on its externo-ventral surface there is a small convex articular face for the diminutive fifth metatarsal.

The number of tarsalia present in the distal row in the other ceratopsians is questionable. The usual number preserved is two, and occasionally there are three. As mentioned earlier, in the mounted skeleton of *Triceratops* in the American Museum (No. 5033) the large, and only tarsal preserved is the calcaneum. In *Monoclonius*, two in the distal row are preserved, one of which was found in place. The largest of these articulated with metatarsal IV. Lull, in his description of *Monoclonius (Centrosaurus) flexus* in the Yale Peabody Museum (No. 2015), states, "Of the distal tarsalia the larger, according to Brown, articulated with the fourth metatarsal. The position of the other is doubtful; possibly it lay above metatarsal III" (1933: 62). Yet in his figure of the pes of this specimen (fig. 29) Lull has shown the largest of these tarsalia above metatarsals I and II, and none above metatarsal IV. In *Chasmosaurus*, according to Lull's discussion (1933: 69) of the two fine skeletons of *C. belli* on exhibition in the National Museum of Canada, there are three tarsalia in the distal row. Their position, however, seems questionable since Sternberg, in his description of these skeletons, says, "In neither specimen was a com-

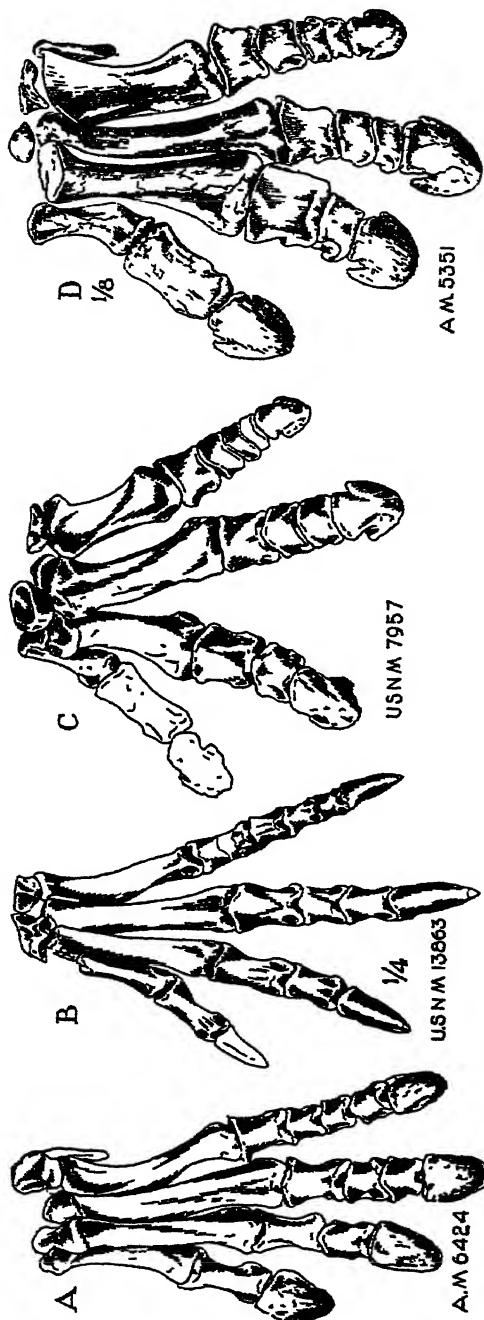


FIGURE 38. Comparative drawings of the left hind foot of four ceratopians. A, *Protoceratops andrewsi*, a fully adult individual. B, *Leptaceratops* sp. (reversed). Modified from Gilmore. C, *Brachyceratops montensis*. Modified from Gilmore. D, *Monoclonius (Centrosaurus) nasicornis*.

plete fore or hind foot preserved, but from what bones are present there seems to have been little difference between those elements in *Chasmosaurus* and *Centrosaurus (Monoclonius)*" (1927: 72). No mention of the size or position of these tarsalia is made by either Lull or Sternberg. In describing the skeleton of *Brachyceratops*, Gilmore states (1917: 33) that three tarsalia are present in the distal row of the tarsus, and says, "The largest articulates with the proximal end of metatarsal II, the smallest with metatarsal III (see fig. 46), to which it was found securely attached by matrix. The third tarsal was wholly in apposition to metatarsal IV." Gilmore has also described a tarsus of *Leptoceratops* (1939: 7). In this specimen (U.S.N.M. 13863) two tarsalia were preserved in the distal row. Although their exact relationship to the metatarsals was not definite, the position Gilmore assigned to them, as shown in figures seven and eight of his paper, seems correct.

The presence of three tarsalia in the distal row of the tarsus in *Brachyceratops* and *Chasmosaurus* may be only a variable feature, but strongly suggests that three were present in all of the later forms. This increase in number is expectant in these forms in which the pes becomes less compact, and becomes relatively shorter and more massive.

In *Protoceratops*, the length of each metatarsal in proportion to the others is about the same as in all other ceratopsians. The metatarsus of *Protoceratops*, however, does possess several unique features. It is very much more compact than in any other form. The proximal extremities of the metatarsals are, therefore, very closely applied to one another, and the first three are especially narrow transversely, and expanded antero-posteriorly. In proportion to the other metatarsals, the third is larger than in any other form, and the whole metatarsus is proportionately longer and more slender. In the later ceratopsians, the metatarsus becomes very much broadened and relatively shortened. Each metatarsal becomes relatively more robust and transversely expanded.

The phalangeal formula is 2, 3, 4, 5, 0. This formula is constant throughout the Ceratopsia. The phalanges are typical for the group, except that the first of digit one is relatively shorter, and all are proportionately more slender and more elongated. The ungual phalanges are similar to those of the manus, except that they are very much larger. Their postero-lateral margins are also perforated instead of being notched as in the later forms. They show the same tendency towards broadening and flattening which is so characteristic of the

more progressive forms. In their general features, therefore, they are decidedly more advanced than are the more claw-like ungual phalanges of *Leptoceratops*. As pointed out above, the same is true for the manus. This is the only important feature of the feet, however, in which *Leptoceratops* is the more primitive.

In all of its unique features, such as, its proportionately large size, its pronounced elongation, and its very compact metatarsus, the hind foot of *Protoceratops* is primitive,—that is, primitive in so far as the ceratopsians are concerned. These are simply heritage characters from an earlier bipedal, ornithischian ancestral stage. In the later forms, as shown in part by *Leptoceratops*, and by *Brachyceratops* and *Monoclonius*, the whole foot broadens and shortens, and the metatarsus becomes less and less compact (see FIGURE 33). Because of these facts, Gilmore's observation that the pes of *Leptoceratops* is ". . . one of the most specialized hind feet of any known quadrupedal dinosaur with the exception of *Protoceratops* . . ." (1939: 11) cannot, therefore, be correct.

MEASUREMENTS OF THE APPENDICULAR SKELETON

	Am. Mus. No. 6419	Am. Mus. No. 6471	Am. Mus. No. 6424
	mm.	mm.	mm.
Length of scapula.....	64	198	231
Width of lower end of scapula.....	16	50	70
Width of upper end of scapula.....	13	47	58
Greatest width of coracoid.....	22.5	62	...
Greatest depth of coracoid.....	21.5	57	...
Length of clavicle.....	...	52	...
Length of humerus.....	55	152	220*
Width of proximal end of humerus.....	13	50	67
Width of distal end of humerus.....	12	43	...
Length of ulna.....	...	126	152*
Length of radius.....	42	112	135
Length of metacarpal I.....	31
Length of metacarpal II.....	41
Length of metacarpal III.....	40
Length of metacarpal IV.....	28.5
Length of metacarpal V.....	15.5
Length of phalanx I ¹	17
Length of phalanx I ²	20
Length of phalanx II ¹	27
Length of phalanx II ²	18.5
Length of phalanx II ³	19.5
Length of phalanx III ¹	15.5

* Estimated.

	Am. Mus. No. 6419	Am. Mus. No. 6471	Am. Mus. No. 6424
	mm.	mm.	mm.
Length of phalanx III ²	13
Length of phalanx III ³	11.5
Length of phalanx III ⁴	15.5
Length of phalanx IV ¹	12
Length of phalanx IV ²	7.5
Length of phalanx IV ³	4.5
Length of phalanx V ¹	10
Length of ilium.	224	240	
Depth of ilium at ischiac peduncle.	58	76	
Length of ischium.	245	330	
Length of femur from top of "greater trochanter" to bottom of external condyle.	187	247	
Greatest width of distal end of femur.	36	65	
Length of tibia.	205	273	
Greatest width of proximal end of tibia.	64	67	
Length of tibia including astragalus.	...	278	
Greatest width of distal end of tibia.	45*	82	
Length of fibula.	193	256	
Length of fibula including calcaneum.	...	279	
Length of metatarsal I.	...	82	
Length of metatarsal II.	...	116	
Length of metatarsal III.	...	125	
Length of metatarsal IV.	...	103	
Length of metatarsal V (estimated from Am. Mus. No. 6478).	...	35	
Length of phalanx I ¹	42	
Length of phalanx I ²	35	
Length of phalanx II ¹	34	
Length of phalanx II ²	29	
Length of phalanx II ³	38	
Length of phalanx III ¹	32	
Length of phalanx III ²	24	
Length of phalanx III ³	23	
Length of phalanx III ⁴	35	
Length of phalanx IV ¹	28	
Length of phalanx IV ²	21	
Length of phalanx IV ³	19	
Length of phalanx IV ⁴	24	
Length of phalanx IV ⁵	33	

MEASUREMENTS OF AN OLD INDIVIDUAL AM. MUS. NO. 6466

mm.

Length of ilium.	345
Depth of ilium at ischiac peduncle.	100
Length of posteriorly deflected portion of pubis from anterior margin of obturator foramen.	118
Total length of pubis.	174
Length of ischium.	380*

* Estimated.

SUMMARY OF THE SALIENT GROWTH CHANGES IN THE POST-CRANIAL SKELETON

Unfortunately in the collection of *Protoceratops*, skeletal material representing the various growth stages is not as abundant as is the material of the skull and lower jaws. Stages from the young adult to the adult are represented, however, by more than a dozen complete or partially complete skeletons. Although, the very immature individual is represented by only two specimens. The most complete of these is Am. Mus. No. 6419, of which the pectoral girdle, and the proximal elements of the fore limbs are in an excellent state of preservation.

In the above study of the post-cranial skeleton, it was shown that some modification takes place in all of the elements from the young to the adult stage. As in the case of the skull and lower jaws, however, certain elements display these modifications more strikingly than others. The more important of these growth changes are the following:

Young

1. Backward slant of neural spines and arches of the dorsals, especially in the median area of the series, very pronounced.
2. Twelfth dorsal completely distinct from the first sacral.
3. Zygapophyses of the sacrals only partially coössified. Distal ends of parapophysial ribs only slightly coössified, and acetabular bar only slightly arched. Number of sacrals whose centra are coössified is four.
4. Eleventh dorsal rib long, free, and in front of the anterior end of the ilium.
5. Scapula narrow and thin, especially just above glenoid cavity. Upper border of glenoid cavity not very heavy and does not extend very much posteriorly. Upper end does not curve forward and anterior margin straight. Ridge on outer surface extends obliquely across that surface. Viewed from the front, scapula quite straight.

Adult

1. Backward slant of neural spines and arches of the dorsals, especially in the median area of the series, less pronounced.
2. Twelfth dorsal partially coössified with the first sacral.
3. Zygapophyses of the sacrals quite completely fused. Distal end of parapophysial ribs fused, and acetabular bar quite arched. Number of sacrals whose centra are coössified is eight.
4. Eleventh dorsal rib shortened, and against and under the antero-inferior surface of the ilium.
5. Scapula broader and thicker, especially above glenoid cavity. Upper border of glenoid cavity heavier and more extended posteriorly. Upper end curves forward and anterior margin rather deeply concave. Ridge on outer surface occupies a more central position. Viewed from the front, scapula pronouncedly convex outward.

Young

6. Humerus slender, extremities narrow, and delto-pectoral crest proximal in position.
7. Ulna and radius slender and extremities relatively unexpanded. Proximal articular surface of radius faces more upward than forward.
8. Anterior portion of ilium slightly deflected outward, its ventral shelf narrow, and the anterior portion of its dorsal margin turned outward only slightly. Ischiac peduncle small.
9. Prepubis short and its distal end slightly expanded dorso-ventrally.
10. Femur slender, "lesser trochanter" distinctly marked off from the "greater trochanter," and lower limit of the internal condyle far above that of the external condyle.

Adult

6. Humerus tends to become heavier, extremities wider, and its delto-pectoral crest more distal in position.
7. Ulna and radius less slender and extremities more expanded. Proximal articular surface of radius faces more forward.
8. Region of ilium anterior to post-acetabular portion increases in length. Anterior portion more deflected outward, its ventral shelf broadens, and the anterior portion of its dorsal margin more outwardly turned. Ischiac peduncle large, and depth of ilium at this peduncle increases.
9. Prepubis somewhat longer and its distal end more expanded dorso-ventrally.
10. Femur somewhat more robust, "lesser trochanter" more closely affiliated with the "greater trochanter," and lower limit of internal condyle more nearly approximates that of the external condyle.

SUMMARY OF THE PRIMITIVE CHARACTERS OF THE POST-CRANIAL SKELETON

Every element in the post-cranial skeleton substantiates the evidence so abundantly shown by the skull and lower jaws, that *Protoceratops* is in nearly all of its characters the most archaic of the known ceratopsians. Throughout our comparative study of the post-cranial skeleton, the procedure in evaluating primitive characters was the same as that employed in the study of the skull and lower jaws. The most important of these primitive characters are here summarized.

1. Hypocentrum (intercentrum) in front of the atlas small, and dorso-lateral projections low. Centrum of atlas shallowly concave anteriorly.
2. Neural arches of the first three cervicals high, only partially coössified, and their intervertebral foramina large.
3. Capitular facets on the atlas well formed.
4. Neural spine of the axis erect and hatchet-shaped.

5. Capitular facet on most of the dorsals at the base of the transverse process, and neural spines and arches, especially on the mid-dorsals, have a pronounced backward slant. Ends of centra round in cross-section.

6. Number of sacrals eight in the adult. Sacrum long, narrow, little arched, sides straight, and greatest width is across the parapophyseal ribs of the first "true" sacral. Median ventral groove on centra of sacrals absent, and neural spines high and not ossified.

7. Midcaudal vertebrae with extremely tall and erect spines; the transverse processes located high on the centra; and, centra never deeper and wider than long.

8. Sternal plates long and narrow.

9. Scapula long and slender; ventral margin faces downward and backward with respect to the long axis; ridge on outer surface oblique; and, less than one-half of glenoid cavity formed by scapula.

10. Coracoid deep, its ventral border markedly curved, and its ventral posterior projection distinctly pointed. Forms more than half of the glenoid cavity.

11. Fore limb relatively small in size—being scarcely more than one-half the length of the hind limb.

12. Humerus elongated and slender, its extremities relatively narrow, and the delto-pectoral crest proximal in position.

13. Ulna and radius relatively slender and their extremities not expanded.

14. Manus relatively small. Metacarpal IV only about three-fourths the length of I, and V only one-half the size of I. Phalanges slender. Unguals elongated and pointed, not subequal in size, and with postero-lateral margins pierced by foramina—not grooved or notched.

15. Ilium with erect dorsal margin.

16. Prepubis small. Acetabular portion of pubis deep and heavy. Posterior process originates anteriorly, has its basal portion markedly curved inward, and is unusually long. Obturator foramen large.

17. Ischium relatively, long, slender, and straight. Its distal end still somewhat dorso-ventrally expanded.

18. Hind limb relatively long.

19. Femur shorter than the tibia, and relatively slender. Head small, faces inwardly, well differentiated from the shaft, and above the level of the "greater trochanter." Fourth trochanter pendant. Distal extremity well above ventral limit of the external condyle.

20. Tibia long and slender, and proximal end but slightly expanded.
21. Pes relatively large and elongate. Astragalus large and deep. Calcaneum deep, free from the astragalus, and distal articular surface large. Two tarsalia in the distal row. Metatarsus very compact. Phalanges slender and elongate. The first of digit one is relatively short. Unguals elongate and pointed, and their postero-lateral margins perforated instead of notched.

ONTOGENETIC CHARACTERS OF *PROTOCERATOPS* THAT FORESHADOW THE EVOLUTION OF THE LATER CERATOPSIANS

An analysis of the primitive characters of *Protoceratops* shows that these characters are more numerous and have greater emphasis in the very immature individual. In its whole form and structure, therefore, the young skeleton recalls many of the features which must have been dominantly characteristic of the earlier ceratopsian ancestor. In its development from the young to the adult, however, the skeleton more closely approximates that of the later ceratopsians, and it is a remarkable fact that during this development the main evolutionary trends that take place in the later forms are forecast.

The most important of these characters that foreshadow the evolutionary trends in the later ceratopsians are the following:

1. Nasal elongates, narrows above, and arches upward to form an incipient horn-core. It becomes more deeply notched at the superior border of the narial opening, which grows proportionately larger and assumes a more upright position.
2. Frill proportionately elongated and widened.
3. A well developed parieto-frontal depression appears which shows the beginning of secondary roofing of the skull.
4. Frontal becomes relatively reduced and its exposure on the orbital border restricted.
5. Postorbital grows forward, arches upward, and develops a very rugose surface, suggesting the first stage in brow horn-core development.
6. Orbita proportionately smaller.
7. Width of the anterior branch of the premaxillary relatively broader.
8. Lachrymal proportionately reduced and the anterior border becomes more erect.
9. Prefrontal grows back to form more of the superior orbital border.

10. Antero-posterior branch of the squamosal becomes relatively larger and inclined upward anteriorly.
11. Lateral temporal opening somewhat reduced.
12. Jugal assumes a more vertical position, and a more extensive contact with the lachrymal.
13. Area of ectopterygoid exposed on palate reduced, showing a tendency for its elimination from the palate which takes place in the later forms.
14. Exoccipitals enter into the composition of the condyle.
15. Neural spines and arches of the dorsals, especially in the median area of the series, become less directed backward.
16. Number of sacrals with coössified centra increases from four to eight, and zygapophyses, and distal ends of parapophysial ribs become quite completely fused. The acetabular bar becomes heavier and more arched.
17. Scapula becomes broader and thicker, especially above the glenoid cavity; upper end curves forward and the anterior margin becomes concave; and, the ridge on the outer surface occupies a more central position.
18. Humerus tends to become relatively more robust, its extremities widen, and the delto-pectoral crest shifts to a more distal position.
19. Ulna and radius become relatively stouter, their extremities become relatively wider, and the proximal articular surface of the radius faces more forward.
20. There is a proportionate increase in length of the ilium in front of the postacetabular portion. The anterior portion becomes more outwardly deflected, its ventral shelf broadens, and the anterior part of its dorsal margin becomes more outwardly turned. The ischiac peduncle enlarges, and the depth of the ilium at this peduncle increases.
21. The prepubis elongates and its distal end expands more dorsoventrally.
22. The femur becomes somewhat more robust, the "lesser trochanter" more closely associated with the "greater trochanter," and the lower limit of the internal condyle more nearly approaches that of the external condyle.

THE EGGS OF PROTOCERATOPS

Of all the important discoveries made by the various American Museum Central Asiatic Expeditions, none has had such widespread popular attention as the discovery of the first eggs in unquestionable association with dinosaur remains.

The first egg was found in 1922 by Dr. Walter Granger, who has related most interestingly the facts concerning this discovery in a brief article published in *Natural History* (1936: 21-25). The rest of the specimens were collected by later expeditions to the Djadochta locality in 1923 and 1925. The entire collection consists of over fifty more or less complete eggs and thousands of shell fragments. A detailed list of the specimens is given in the appendix. They were found as individual eggs weathered out and lying on the surface, and as remnants of nests. Two of the nests are quite complete. One (Am. Mus. No. 6508) is a group of fifteen more or less complete eggs, thirteen of which are *in situ*, and fragments of probably two others that are weathered out and broken up (see PLATE 12). These eggs represent approximately one-half of the original nest—the other half being weathered away. The entire nest probably consisted of thirty or thirty-five eggs. They apparently were deposited in circular fashion, and occur in three layers—those deposited first being in the lower part of the nest, and those last in the upper part. A significant feature of this nest is that all of the eggs in the upper region were broken before burial, and the skeleton of *Oviraptor* (Am. Mus. No. 6517) was found lying directly over the nest with only four inches of matrix intervening.

The other rather complete nest is Am. Mus. No. 6631, which consists of eighteen eggs in their original position, but with the tops of all of them sheared off by weathering. The arrangement of the eggs is circular. In the center is a group of five eggs that were deposited in the deepest portion of the nest. Outside of this group, and at a higher level, are eleven more eggs forming a circle around the inner group. Only the lower ends of two other eggs of the outer and still higher circle are preserved (see PLATE 12). The probable methods used by the female *Protoceratops* in constructing her nest and depositing her eggs in this orderly arrangement has been suggested by Dr. Granger (1936: 23). Also, his suggestion that sudden burial by drifting sands prevented the eggs in this nest from hatching seems very logical.

The variation in size, form, and surface texture is very pronounced. The smallest egg is between three and four inches long and the largest is nearly eight inches long.* The variation in the diameter is relatively not as great. In form, the eggs are long ovate with one end always

* A very small egg (Am. Mus. No. 6654) not quite an inch long, and with the shell entirely gone—leaving only the internal mold, was also recovered from the Djadochta beds. Its general form is oval, and in this feature compares so favorably with that of the present-day crocodilian eggs, that it seems reasonable to assume that it probably is an egg of *Shamosuchus*—a small crocodilian from the same beds.

smaller than the other. Some, however, are less pointed than others. This holds true for both the small and the large types of eggs. Variation in surface texture is quite marked. The surfaces of the smaller eggs always appear to be smooth, while those of the larger eggs are vermiculated. The vermiculations on some are not as pronounced as on others, and on some they become quite node-like instead of elongated. These surface features are never as pronounced on the ends as on the rest of the egg, although there is a considerable amount of variation in this respect. When seen under the binoculars the smaller, and smoother eggs, present the same patterns of vermiculations as in the larger eggs. They are very much less developed, however, and the eggs appear quite smooth. It would seem, therefore, that pronounced surface texture is correlated with large size. There also seems to be a correlation between size and the thickness of the shell. In the larger, and especially rugose eggs, the shell seems slightly thicker.

Because of the marked variation in size, and the appreciable variation in form and in surface texture, the question of whether or not all of the eggs in the Djadochta collection (excluding the crocodilian egg mentioned above) can be assigned to *Protoceratops*, immediately arises. Van Straelen (1925: 1-3) made a study of the micro-structure of samples from three different groups of eggs in the collection. Two samples were from the nest of fifteen eggs (Am. Mus. No. 6508) described above, which represent the large type of egg with the somewhat node-like vermiculations and with the rather blunt smaller end. Another sample was from a group of five smaller eggs whose surfaces are definitely without the strong vermiculations (Am. Mus. 6511). The third was a group of three large eggs representing the markedly vermiculated type with the rather pointed smaller end. According to Van Straelen, the micro-structure is the same in all of these specimens. In a subsequent paper, however (1928 [1927]: 307), he suggests that two species are represented, and states that the second differs from those which were assigned to *Protoceratops* in his first paper in that the shell is thicker and the vermiculations are more pronounced and farther apart. Concerning the micro-structure of this type, he goes on to say, "La structure microscopique est identique à celle des premiers œufs décrits, sauf en ce que concerne les canaux aérifères encore plus fins et plus rares." He in no way designates which specimens of the collection he used for this determination. These distinctions do not seem to justify his conclusion that two distinct forms are represented. Thickness of shell seems correlated with size, and there seems to be no constancy of pattern of vermiculations.

Furthermore, taking into consideration the marked variation in the eggs of some of the present day reptiles, the suggestion of more than one form being represented by the Djadochta eggs is even less warranted. For example, Reese reports (1923-8) a maximum variation in length in caiman eggs of more than fifty per cent of the length of the shortest egg, and illustrates (figure 4) a variation in shape that is as great as that found in *Protoceratops*. He further states that in surface texture some eggs are comparatively smooth, and in others, ". . . the surface of the shell is extremely harsh." Variations in the eggs of some of the other crocodilians are equally as pronounced.

At present there seems to be no good evidence for contending that any of the Djadochta eggs* are from a form other than *Protoceratops*. In a forthcoming paper, the problem of the Djadochta eggs will be thoroughly treated. Comparisons in variations of form and structure will be made with the eggs of recent reptiles. A study of the micro-structure of samples from all of the nests will be presented, and two specimens which show convincingly the presence of well-developed embryos will be thoroughly described.

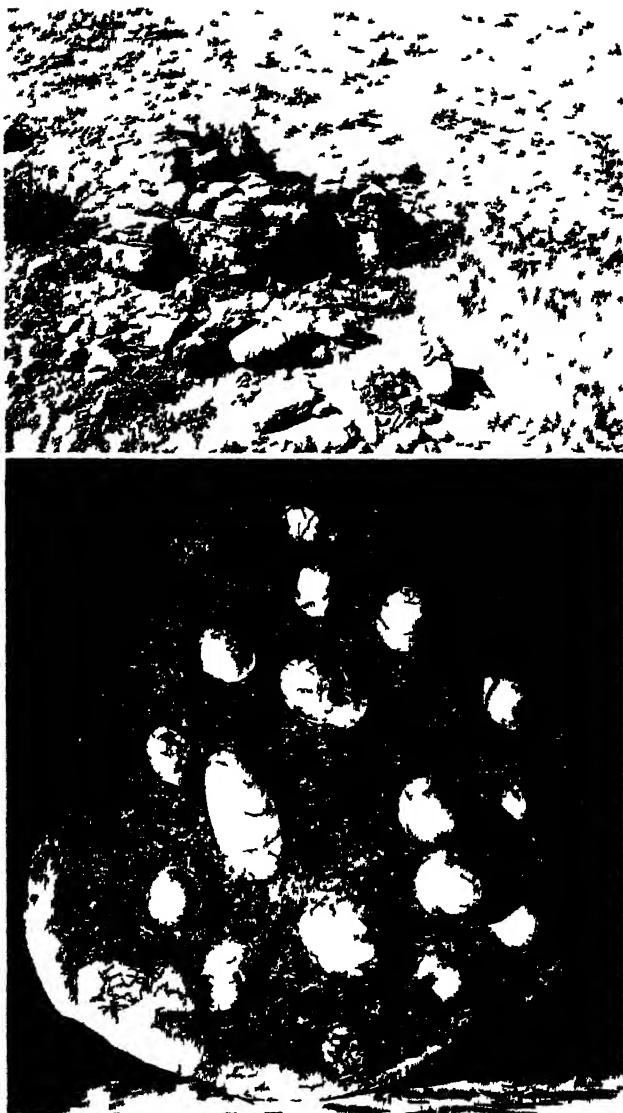
THE INTEGUMENT OF PROTOCERATOPS

Only one specimen (Am. Mus. No. 6418) presents any suggestion of the integument. This is a nearly complete skeleton, which was found in an unusual curled up position. A thin, hard, and wrinkled layer of matrix covers a considerable portion of the skull and jaws. The wrinkling has a very skin-like appearance, and is most predominant on the left side of the head over the orbit, at the corner of the mouth just in front of the jugal, and over the side of the frill and lateral temporal opening (see PLATE 13). As far as can be determined, it is without any trace of skin-structure. The surface form of the original skin, however, undoubtedly influenced the form of the matrix at burial, and by the chemical action set up through its decay probably caused this thin layer of matrix surrounding the bone to become considerably indurated.

THE FAMILY PROTOCERATOPSIDAE

From the detailed comparative study given in the preceding pages, it becomes apparent that *Protoceratops*, in all of its salient features, is an

* Except the probable crocodilian egg mentioned above, and a few fragments (Am. Mus. No. 6660) whose surfaces present vermiculations quite distinct from the other specimens. These fragments come from the extreme western end of the Djadochta exposures and were not found in association with *Protoceratops*.



Two nests of *Protoceratops andrewsi*. Above, the first nest discovered. Fifteen eggs were found in position under the sandstone. Approximately one half of the nest had been destroyed by erosion although the two eggs in the foreground resisted weathering. Below, under side of a nest with eighteen eggs. Note their circular arrangement.

BROWN AND SCHLAIKJER PROTOCERATOPS



Protoceratops andrewsi. Two views of a skeleton in an unusual curled up position showing a thin, indurated layer of matrix over the head region which has a very skin like appearance. Am. Mus. No. 6418, a young adult individual.

ideal progenitor for the later ceratopsians. This ancestral position, however, may be only structural because of the possible lateness in time of the Djadochta beds in which the remains of this form occur. Nevertheless, the impressive array of outstanding primitive characters summarized in the preceding account makes its position far down the scale of ceratopsian evolution indisputable. Throughout this study it has also been shown that *Leptoceratops*, from the Edmonton Cretaceous, is but a slightly more progressive form. Because of the close affinity between these two genera, and because they are distinct from the more progressive ceratopsians in such a large number of outstanding characters, they should be placed in a separate family, the Protoceratopsidae. It should be emphasized again, however, that all of these outstanding unique characters are primitive characters, and that *Protoceratops* and *Leptoceratops* are distinct from the known members of the Ceratopsidae because they are primitive.

The family name, Protoceratopsidae, was first suggested by Granger and Gregory (1923: 4). They designated *Protoceratops* as the type of this new family and characterized it, ". . . by the lack of horns, the very large size of the orbits, and the narrowness of the postorbital-squamosal bar." In a brief subsequent paper Gregory and Mook (1925: 4) included *Leptoceratops* in the Protoceratopsidae and gave the following extended definition of this family:

"Primitive small ceratopsians, with a hornless skull, without either secondary skull roof or pseudopineal foramen above the frontals, no epoccipital bones; with simple oval anterior nares and unspecialized premaxillae. A well-developed occipital frill, with large transversely oval parietal fontanelles. Freely articulating palpebral bones (supraorbitals) attached to the anterosuperior corner of the orbits. Premaxillaries with teeth. Cheek teeth arranged in a vertical series of not more than two developed at one time; roots simple (not bifid). Fore limb slender, manus much smaller and shorter than pes, the latter elongate, compressed. Sacral complex of seven or eight vertebrae. Ilium with blade but slightly inclined outward to the sagittal plane, not reflected or produced laterally above the femur. Prepubic process relatively small, not expanded vertically; postpubic process but little reduced. Femur with large fourth trochanter, femur shorter than tibia. Midcaudal vertebrae with very long spines."

As the result of our study, it now becomes necessary to modify and extend this definition. Since the old "male" skulls of *Protoceratops* show an early stage in the development of the ceratopsian nasal horn-core, the designation of a "hornless skull" for this family is not entirely warranted. Furthermore, the lack of "secondary skull roof or pseudopineal foramen above the frontals," and absence of epoccipital bones are not distinctive features. *Brachyceratops* has no secondary

skull roof, and the presence of a pseudopineal foramen in any ceratopsian, especially in the earlier more advanced forms, is questionable. As for epoccipital bones, they were not present in *Brachyceratops*, and probably were absent in some of the end members of the group. Likewise, "parietal fontanelles" in the frill are so variable in size and form, even in *Protoceratops*, that they are of no particular taxonomic significance. The remainder of the above characteristics are sound, but important additions are necessary. We, therefore, propose the following definition of the family:

1. Narial opening small and no well-developed fossa in front of it.
2. Premaxillary proportionately large, and top of posterior branch even with, or above the dorsal margin of the lachrymal. Alveoli of two teeth present, and no well-formed septum.
3. Preorbital fossa large.
4. Nasal relatively small, and in the adult "male" it becomes arched to form an incipient horn-core.
5. Orbit proportionately large.
6. Prefrontals do not meet in the midline, and they form part of orbital border.
7. Palpebrals freely articulate with the prefrontals, and are not, therefore, taken over into the skull roof to form the "supraorbitals." Not in contact with the lachrymal and postorbitals.
8. Frontal in contact with the nasal, and forms part of the orbital border.
9. Postorbital presents only the first suggestion of a brow horn-core in that it is arched and rugose in the adult. Together with the squamosal it forms the primitive narrow postorbital-squamosal bar.
10. Squamosal small, and does not unite with the jugal and (or) quadratojugal behind the lateral temporal opening.
11. Lateral temporal opening large.
12. Frill short and has a high median crest. Entire structure acted as an anchor for the large capiti-mandibularis muscle masses.
13. Quadrato and exoccipital do not unite—the posterior ventral projection of the squamosal intervenes.
14. Exoccipitals do not unite above the foramen magnum.
15. Ectopterygoid large, extensively exposed on the palate, and is in contact with the jugal.
16. Dentary short and deep, and ventral border curved downward. Coronoid process low, set close to tooth-row, and dorsal end not expanded.
17. Teeth single-rooted. Not more than fifteen vertical series of

teeth, and only two or possibly three teeth in each. No tube-like casing of spongy bone for each vertical series.

18. Neural arches of the first three cervicals high, only partially coössified, and their intervertebral foramina large. Neural spine of axis erect and hatchet-shaped.

19. Number of sacrals eight in the adult. Sacrum long, narrow, little arched, and neural spines high and not coössified.

20. Midcaudal vertebrae with extremely tall and erect spines.

21. Fore limb relatively small in size.

22. Unguals elongated and pointed.

23. Ilium with erect dorsal margin.

24. Pubis small. Prepubic process short. Obturator foramen large.

25. Ischium long, slender, and comparatively straight.

26. Hind limb relatively long and slender. Femur shorter than tibia. Pes relatively large and elongated, metatarsus compact.

APPENDIX

Protoceratops Specimens in The American Museum of Natural History

AM. MUS.	FIELD No.	DESCRIPTION OF SPECIMEN	YEAR COLLECTED
6251	102	Skull, <i>Type</i> .	1922
6273	102	Skull and jaws, small.	1922
6274	102	Miscell. jaw fragments and foot-bones.	1922
6408	372	Skull and jaws.	1923
6409	295	Skull, nearly perfect.	1923
6413	324	Skull and jaws.	1923
6414	7356	Skull and jaws, large.	1923
6416	373	Skull, jaws and skeleton, complete to proximal caudals.	1923
6417	309	Skull, jaws and nearly complete skeleton. Mounted.	1923
6418	265	Skull, jaws and most of skeleton to middle of tail.	1923
6419	366	Skull, jaws, portions of anterior part of skeleton, small.	1923
6421	259	Skull and jaws in concretion, small.	1923
6422	288	Fragmentary skull and jaws, some skeletal fragments, small.	1923
6423	296	Skull and jaws in concretion, weathered.	1923
6424	274	Large part of skeleton, without skull and jaws.	1923
6425	349	Skull and jaws, nearly perfect, large.	1923
6426	335	Skull, jaws, part of skeleton.	1923
6428	371	Skull and jaws.	1923
6429	273	Skull, no teeth.	1923
6430	315	Skull and jaws.	1923

AM. MUS.	FIELD No.	DESCRIPTION OF SPECIMEN	YEAR COLLECTED
6431	315	Skull and jaws.	1923
6432	374	Skull and jaws.	1923
6433	353	Skull and jaws, fine.	1923
6434	375	Skull and jaws.	1923
6436	—	Teeth.	1923
6437	297	Skull and jaws in concretion.	1923
6438	333	Skull and jaws, large.	1923
6439	350	Skull and jaws.	1923
6440	305	Skull and jaws, panel mount.	1923
6441	369	Skull and jaws, panel mount.	1923
6442	368	Skull, imperfect.	1923
6443	319	Skull and jaws, crushed.	1923
6444	355	Skull and jaws.	1923
6445	344	Skull fragmentary in matrix.	1923
6446	270	Skull fragmentary in matrix.	1923
6447	365	Skull and jaws.	1923
6448	312	Skull and jaws.	1923
6449	317	Skull fragmentary.	1923
6450	318	Skull, jaws, part of skeleton.	1923
6451	370	Fragmentary skull and part of skeleton.	1923
6452	302	Skull, part of skeleton.	1923
6453	272	Fragmentary skeleton.	1923
6454	299	Weathered skull and much of skeleton.	1923
6458	323	Skull fragmentary.	1923
6459	255	Lower jaw.	1923
6460	—	Lower jaws.	1923
6461	266	Skull and jaws.	1923
6463	352	Lower jaws.	1923
6465	347	Skull and partial skeleton.	1923
6466	513	Large skull and jaws, and thoracic section.	1925
6467	514	Skull, jaws, and nearly complete skeleton, mounted.	1925
6468	558	Skull fragmentary.	1925
6469	354	Skull fragments and parts of skeleton.	1923
6470	316	Back of skull and parts of skeleton.	1923
6471	515	Skull and jaws with part of skeleton.	1925
6472	556	Incomplete skull and jaws.	1925
6473	557	Part of skull and jaws, poor.	1925
6474	511	Caudal vertebrae and hind limb.	1925
6475	331	Skeletal parts.	1923
6477	326	Imperfect skull.	1923
6478	320	Caudals, ulna, radius, fore and hind foot.	1923
6479	345	Partial skeleton.	1923
6480	341	Partial skull and jaws in weathered concretion.	1923
6481	342	Pelvis.	1923
6482	291	Fore and hind foot.	1923
6483	293	Weathered skull and jaws, fragmentary and small.	1923
6484	292	Skeletal parts.	1923

AM. MUS.	FIELD No.	DESCRIPTION OF SPECIMEN	YEAR COLLECTED
6485	339	Skull in weathered concretion.	1923
6486	310	Weathered skull and jaws.	1923
6487	306	Weathered skull and jaws.	1923
6488	338	Fragmentary skull in weathered concretion.	1923
6489	327	Skull and jaws.	1923
6490	314	Part of skeleton.	1923
6491	308	Skull and jaws.	1923
6492	261	Fragmentary skeleton.	1923
6493	313	Vertebral column.	1923
6494	348	Pelvis, hind limb and series of vertebrae.	1923
6495	255	Front of skull and jaws.	1923
6496	255	Front of skull and jaws.	1923
6497	560	Fragment of lower jaw.	1925
6498	—	Left dentary, very young.	1923
6499	—	Right dentary, unborn.	1923
6636	294	Poor skull and part of skeleton.	1923
6637	346	Part of a skull.	1923
6638	279	Front of skull.	1923
6639	282	Portion of a skull.	1923
6640	283	Caudal vertebrae.	1923
6644	560	Parts of small skull and jaw, limb-bones, etc.	1925

Protoceratops Specimens Sent to Other Institutions

AM. MUS.	FIELD No.	DESCRIPTION OF SPECIMEN AND WHERE SENT	YEAR COLLECTED
6415	334	Skull and jaws, small, crushed. Sent to Peking, China, 1925.	1923
6420	332	Skull and jaws, largest size. Sent to Field Museum, Chicago, Jan. 1926.	1923
6427	277	Skull and jaws. Sent to Univ. California, 1934.	1923
6435	290	Skull, jaws, most of skeleton. Sent to Field Museum, Chicago, Jan. 1926.	1923
6455	376	Hind foot. Sent to Munich, Germany, 1933.	1923
6456	357	Skull and jaws. Sent to Paris, France, 1927.	1923
6457	322	Half a skull and jaw fragment. Sent to Munich, Germany, 1933.	1923
6462	275	Lower jaw, palate, foot-bones. Sent to Univ. California, 1934.	1923
6464	285	Skull, jaws, part skeleton. Sent to Urga, Mongolia, 1927.	1923
6476	300	Skull and jaws. Sent to Urga, Mongolia, 1927.	1923

AM. MUS.	FIELD No.	DESCRIPTION OF SPECIMEN	YEAR COLLECTED
6641	325	Part of skull and jaws. Sent to Munich, Germany, 1933.	1923
—	269	Front of skull and jaws. Sent to Public School Mus. in Battle Creek, Mich.	1923
—	298	Posterior part of skull in concretion, weathered. Sent to Public School Mus., in Battle Creek, Mich.	1923

The Eggs of *Protoceratops*

AM. MUS.	FIELD No.	DESCRIPTION OF SPECIMEN	YEAR COLLECTED
2905	102	Portion of egg (small).	1922
6505	255	Portions of two eggs.	1923
6506	255	Cast of egg, with few fragments of shell adhering.	1923
6507	255	Finely preserved fragments of one or more eggshells.	1923
6508	267	Group of 15 more or less complete eggs and fragments of probably two others which were weathered out and broken up. Thirteen are in situ. One sent to Colgate University 1924.	1923
6509	351	Group of three weathered eggs. Two showing por- tions of embryos.	1923
6510	287	Group of three weathered eggs.	1923
6511	343	Group of five eggs in matrix.	1923
6512	264	Group of several eggs in matrix, each partly weath- ered off.	1923
6513	281	Group of five eggs in matrix.	1923
6631	526	Group of eighteen eggs, tops sheared off.	1925
6633	562	Eight large eggs, crushed but fairly complete; parts of several others.	1925
6635	—	Group of six small eggs. Sent to the Field Museum, Chicago, 1926.	1925
6642	528	Nest of four small, thin-shelled eggs.	1925
6649	564	Group of badly weathered eggs.	1925
6650	560	Egg.	1925
6651	560	Egg fragments.	1925
6652	560	Eggs—crushed.	1925
6654	560	Very small egg in concretion. Probably crocodilian.	1925
6660	518	Fragments of eggshells from one knoll, probably from one clutch.	1925

BIBLIOGRAPHY

Adams, Leverett Allen, 1877-

1919. A memoir on the phylogeny of the jaw muscles in recent and fossil vertebrates. Ann. N. Y. Acad. Sci. 28: 51-166. figs. 1-5. pls. 1-13. Jan. 15, 1919.

Andrews, Roy Chapman, 1884-

1932. The new conquest of Central Asia. Nat. Hist. Central Asia 1: i-1, 1-678. figs. 1-12. pls. 1-128. 3 folding maps. Dec. 29, 1932.

Berkey, Charles Peter, 1867- ; & Morris, Frederick Kuhne, 1886-

1927. Geology of Mongolia. Nat. Hist. Central Asia 2: i-xxx, 1-475. figs. 1-161. pls. 1-44. 1927.

Brown, Barnum, 1873-

1914. *Anchiceratops*, a new genus of horned dinosaurs from the Edmonton Cretaceous of Alberta. With discussion of the origin of the ceratopsian crest and the brain casts of *Anchiceratops* and *Trachodon*. Bull. Am. Mus. Nat. Hist. 33: 539-548. fig. 1. pls. 29-37. Oct 8, 1914.
1914. *Lepioceratops*, a New Genus of Ceratopsia from the Edmonton Cretaceous of Alberta. Bull. Am. Mus. Nat. Hist. 33: 567-580. pl. 42. Oct. 8, 1914.
1917. A complete skeleton of the horned dinosaur *Monoclonius* and description of a second skeleton showing skin impressions. Bull. Am. Mus. Nat. Hist. 37: 281-306. figs. 1-4. pls. 11-19. May 31, 1917.

; & Schlaikjer, Erich Maren, 1905-

1937. The Skeleton of *Styracosaurus* with the description of a new species. Am. Mus. Novit. No. 955: 1-12, figs. 1-5, Oct. 30, 1937.
- 1940a. The origin of ceratopsian horn cores. Am. Mus. Novit. 1065: 1-8. figs. 1 and 2. May 3, 1940.
- 1940b. A new element in the ceratopsian jaw with additional notes on the mandible. Am. Mus. Novit. in press.

Gilmore, Charles Whitney, 1874-

1917. *Brachyceratops* a ceratopsian dinosaur from the Two Medicine formation of Montana, with notes on associated fossil reptiles. U. S. Geol. Surv. Prof. Pap. 103: 1-45. figs. 1-57. pls. 1-4. 1917.
1919. A new restoration of *Triceratops*, with notes on the osteology of the genus. Proc. U. S. Nat. Mus. 55: 97-112. figs. 1-6. pls. 3-4. 1919.
1930. On dinosaurian reptiles from the Two Medicine formation of Montana. Proc. U. S. Nat. Mus. 77: 1-39. figs. 1-18. pls. 1-10. 1930.
1939. Ceratopsian dinosaurs from the Two Medicine formation, Upper Cretaceous of Montana. Proc. U. S. Nat. Mus. 87: 1-18. figs. 1-11. 1939.

Granger, Walter, 1872- ; & Gregory, William King, 1876-

1923. *Protoceratops andrewsi*, a pre-ceratopsian dinosaur from Mongolia. Am. Mus. Novit. 72: 1-9. figs. 1-4. May 4, 1923.
1936. The story of the dinosaur eggs. Nat. Hist. 38: 21-25. 5 figs. Jan., 1936.

Gregory, William King, 1876-

1920. Studies in comparative myology and osteology. No. IV. A review of the evolution of the lacrymal bone of vertebrates with special reference to that of mammals. Bull. Am. Mus. Nat. Hist. 42: 95-263. figs. 1-196. pls. 1-17. Dec. 4, 1920.
1927. The Mongolian life record. Sci. Monthly 24: 169-181. figs. 1-10. Feb., 1927.

—; & Mook, Charles Craig, 1887-

1925. On *Protoceratops*, a primitive ceratopsian dinosaur from the lower Cretaceous of Mongolia. Am. Mus. Novit. 156: 1-9. figs. 1-3. Feb. 11, 1925.

Hatcher, John Bell, 1861-1904; Marsh, Othniel Charles, 1831-1899; & Lull, Richard Swann, 1867-

1907. The Ceratopsia. Mon. U. S. Geol. Surv. 49: i-xxx, 1-198. figs. 1-120. pls. 1-51. 1907.

Hay, Oliver Perry, 1846-1930

1909. On the skull and brain of *Triceratops*, with notes on the brain-cases of *Iguanodon* and *Megalosaurus*. Proc. U. S. Nat. Mus. 36: 95-108. pls 1-3. 6 F 1909.

Huber, Ernst, 1892-1932

1930. Evolution of facial musculature and cutaneous field of trigeminus. Quart. Rev. Biol. 5: 133-188. figs. 1-26. Jan., 1930.

Huene, Friedrich von,

1911. Beiträge zur Kenntniss des Ceratopsidenschädels. Neues Jahrb. Min. Geol. Pal. 2: 146-162. figs. 1-10. 1911.

Lambe, Lawrence Morris, 1863-1919

1913. A new genus and species of Ceratopsia from the Belly River formation of Alberta. Ottawa Naturalist 27: 109-116. pls. 10-12. Dec., 1913.

Lull, Richard Swann, 1867-

1903. Skull of *Triceratops serratus*. Bull. Am. Mus. Nat. 19: 685-695. fig. 1. pls. 54, 55. Dec. 24, 1903.
1905. Restoration of the horned dinosaur *Diceratops*. Am. Jour. Sci. 20: 420-422. pl. 14. Dec., 1905.
1908. The cranial musculature and the origin of the frill in the ceratopsian dinosaurs. Am. Jour. Sci. 25: 387-399. figs. 1-10. pls. 1-3. May, 1908.
1933. A revision of the Ceratopsia or horned dinosaurs. Mem. Peabody Mus. Nat. Hist. 3 (3): 1-175. figs. 1-42. pls. 1-17. 1933.

Marsh, Othniel Charles, 1831-1899

1816. The dinosaurs of North America. 16th Ann. Rep. U. S. Geol. Surv. 1894-1895, Pt. 1: 133-244. figs. 1-66. pls. 1-85. 1896.

Osborn, Henry Fairfield, 1857-1935

1924. *Petacosaurus* and *Protiguanodon*; two Lower Cretaceous Iguanodonts from Mongolia. Am. Mus. Novit. No. 127: 1-16. figs. 1-9. Sept. 4, 1924.

Reese, Albert Moore, 1872-

1923. Notes on the Crocodilia of British Guiana. Bull. West Va. Univ. Sci. Assoc. 2: 1-12. figs. 1-5. Aug., 1923.

Romer, Alfred Sherwood, 1894-

1922. The locomotor apparatus of certain primitive and mammal-like reptiles. Bull. Am. Mus. Nat. Hist. 46: 517-606. figs. 1-7. pls. 27-46. Oct. 3, 1922.
1927. The pectoral musculature of ornithischian dinosaurs. Acta Zool. 8: 225-276. figs. 1-20. 1927.

Russell, Loris Shano, 1904-

1935. Musculature and function in the Ceratopsia. Bull. Nat. Mus. Canada 77 (geol. ser. 52): 1-48. figs. 1-9.

Schlaikjer, Erich Maren, 1905-

1935. The Torrington Member of the Lance formation and a study of a new *Triceratops*. Bull. Mus. Comp. Zool. 74: 29-68. figs. 1-5. pls. 1-6. Jan., 1935.

Simpson, George Gaylord, 1902-

1937. New reptiles from the Eocene of South America. Am. Mus. Novit. 927: 1-3. May 12, 1937.

Sternberg, Charles Mottram,

1927. Homologies of certain bones of the ceratopsian skull. Trans. Roy. Soc. Canada 31: 135-143. pls. 1-3. 1927.

Van Straelen, Victor Émile, 1889-

1925. The microstructure of the dinosaurian egg-shells from the Cretaceous beds of Mongolia. Am. Mus. Novit. 173: 1-4. figs. 1-2. May 27, 1925.
1928. Les œufs de reptiles fossiles. Palaeobiologica 1: 295-312. pls. 26-28. Dec., [1927].

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

VOLUME XL, ART. 4. PAGES 267-288

DECEMBER 20, 1940

RECURRENT PALEOZOIC CONTINENTAL FACIES IN PENNSYLVANIA*

By

BRADFORD WILLARD†

CONTENTS

	PAGE
INTRODUCTION.....	267
THE CONTINENTAL FACIES.....	268
The Ordovician.....	271
The Silurian.....	273
The Devonian.....	274
The Mississippian.....	275
The Pennsylvanian and Permian.....	277
The Triassic.....	277
THE FACIES SHIFTED.....	278
The Geosyncline Migrated.....	282
THE SHIFT OF THE FACIES SHIFTS.....	284
BIBLIOGRAPHY.....	286

INTRODUCTION

All Paleozoic systems are represented in Pennsylvania. The majority of them are well-developed, marine sequences. Commencing with the Ordovician, there is an occurrence in each system of beds of continental origin. The continental elements, here referred to as facies, alternate with marine formations in the older systems where marine elements are dominant. But in the younger systems, the continental strata dominate to the nearly total exclusion in some of all marine elements. The recurrent continental facies suggest a cyclic

* Publication made possible through a grant from the income of the George Herbert Sherman Memorial Fund.

† Lehigh University.

repetition. The cycles do not always agree in chronology or in stratigraphic boundaries and sequences with the stereotyped periodic and systemic boundaries.

The Paleozoic systems in Pennsylvania are particularly well adapted for the study here outlined because of the dominance of clastic sediments. Toward the close of Trenton time the great limestone depositing seas of the Cambrian and Ordovician disappeared, and terrigenous sediments became the rule. Relatively little Silurian limestone is present. Most of the Devonian is devoid of calcareous strata, and only an occasional thin bed is encountered above the Devonian. Thus, the vast majority of our Paleozoic sediments since Middle Ordovician time are near-shore and probably rather shallow-water deposits. Because these facies are so common, the marine to continental relations and lateral intergradations are present and readily studied from well-preserved sequences.

THE CONTINENTAL FACIES

The presence of continental facies in the older Paleozoic in particular is recorded commonly on the persistent appearance of red beds. These are often traceable into contemporaneously formed, non-red, fossiliferous, marine sediments. However, the color alone is not a sufficiently diagnostic feature, nor is it an infallible criterion. Thus, Cambrian red beds in southeastern Massachusetts carry trilobites and pteropods. Additional, supporting, and partially substantiating evidence is essential. Such corroboration embraces various pseudofossils such as rain-drop or hail imprints, abundant current-formed ripple marks, and to a certain extent, desiccation cracks. Fossils of terrestrial or freshwater organisms are of course valuable supplemental proof of environment at time of origin. To such data must be appended any observation on field relations to subjacent, superjacent and adjacent, or contemporaneously formed beds, particularly both interfacial and intrafacial relations should be stressed.

Red beds may be common and may make up most of a freshwater facies of a particular age. On the other hand, non-red continental beds occur sparingly in the Silurian and are abundant from the Devonian. They dominate the freshwater facies in the Pennsylvanian and Permian. The preponderance of red over non-red beds in the several recurrent facies thus varies roughly directly with the age. The older continental facies recurrents are chiefly red; the younger ones are practically all non-red. The change is progressive and appears to be, superficially at least, a correlate of the development of land floras.

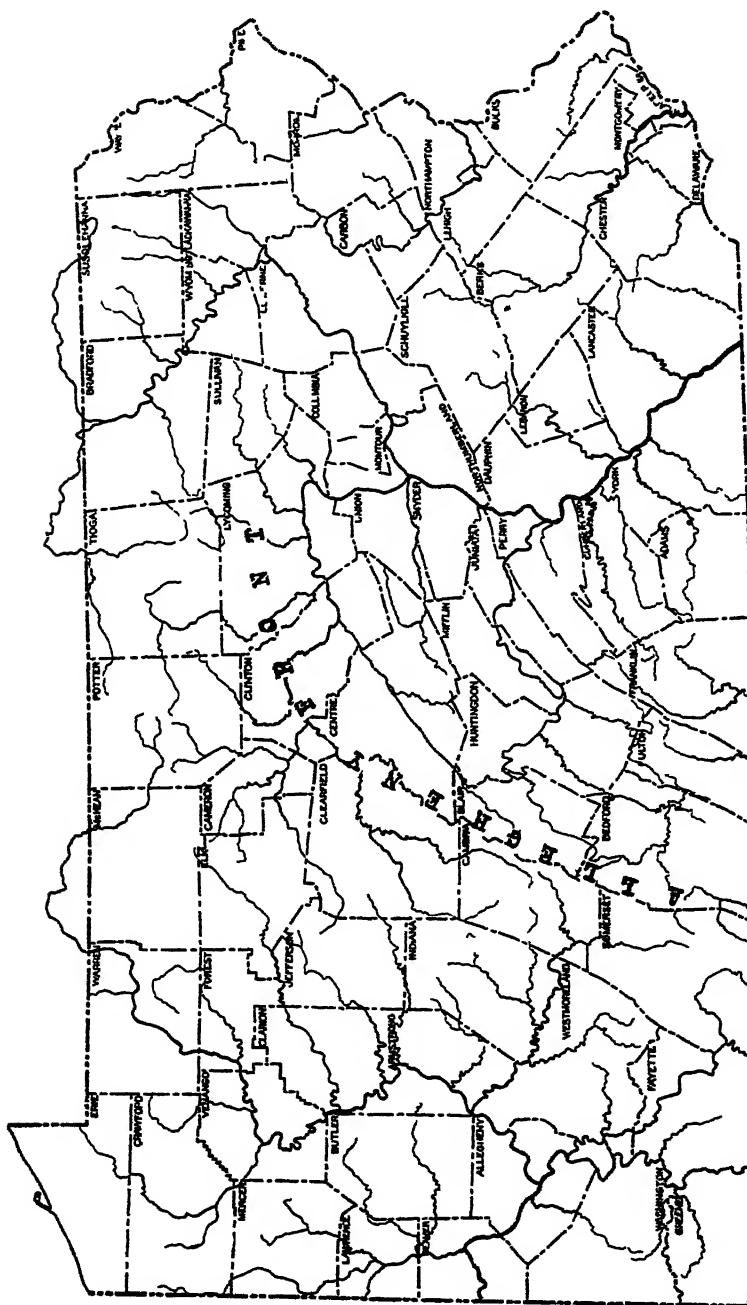


FIGURE 1. Outline map of Pennsylvania.

Nevertheless, there is a less apparent but far more deep-seated and far-reaching cause based directly upon tectonic changes in the area supplying the sediments.

Continental facies have not been identified in the Cambrian of Pennsylvania. They are present in the Ordovician, Silurian, Devonian, Mississippian, Pennsylvanian, and Permian. The great expanse of earlier marine Paleozoic units crops out in eastern, central, and south-central Pennsylvania. It is here that several of the red or gray to greenish continental units are likewise best observable. The Mississippian and Pennsylvanian systems are widely distributed, but the Permian is restricted to the extreme southwest. The area of wide-spread coal measures throughout most of the western half of the State is largely disregarded in the following discussion. In fact, the post-Mississippian may be almost dismissed with the characterization of being gray continental sediments among which red beds and marine intercalations are nearly absent.

Before going further a brief review of certain aspects of the origin of red beds seems to be in order. Many and varied have been the theories advanced for their genesis. Of these, some are totally untenable, others are in part true or applicable in special cases. Probably that of P. E. Raymond is the most reasonable. The writer has applied it with some success in interpreting the origin of the continental Devonian beds in Pennsylvania. It seems equally applicable with some modification and amplifying to the other continental facies as here discussed.

Briefly, the theory would develop our red sediments by reworking a previously formed red regolith. This regolith upon being reworked through ordinary processes of erosion and transportation, is subsequently redeposited as red sandstone, shale, etc. In other words, the theory assumes, in contradistinction to the older belief of Joseph Barrell, that the beds have not been epigenetically reddened where they are laid down, but are syngenetically colored through being derived from already red material. Since we have recurrent red continental facies (omitting for now the gray or green continental members), it is necessary to suppose that from time to time circumstances arose which caused the accumulation of red residual soils. These conditions alternated with intervals during which those soils were removed and redeposited as red strata. Such an alternation of conditions may be attributed to successive intervals of peneplanation and uplift of the land lying to the east and southeast of central and eastern Pennsylvania. Under these circumstances a red regolith could accumulate and

be subsequently eroded and redeposited by streams flowing westward and northwestward off Appalachia.

Among our red beds are certain non-red strata or larger units which are as surely continental as the red with which they alternate. They, too, grade into contemporaneously deposited marine formations, and may carry characteristic fossils and pseudofossils. Returning to the suggestion of alternate quiescence with peneplanation and uplift with accelerated erosion, let us suppose that erosion and transportation have removed substantially all of the red residual soil at some time prior to peneplanation. Erosion continues, but instead of the streams bearing the red mud, they now cut deeper into the unaltered or partly weathered to fresh rock, which may be any color, and which produces non-red sediments that are spread seaward and grade into marine beds as readily as did the red.

The Ordovician

The Ordovician sequence in Pennsylvania is complex. We need consider only part of it, or that portion which includes beds of possible continental origin. In south-central Pennsylvania, if the Juniata is included, the Ordovician sequence is typified by the following (Willard 1939) :

SILURIAN	
ORDOVICIAN...	Juniata (red) formation
	Bald Eagle member at base
	Fairview or Shochary sandstones*
	Martinsburg "formation"
	Jonestown (red) beds
	Unnamed and undifferentiated dark, marine shales, etc.†
	Cocalico shale
	Chambersburg limestone and correlates
	Stones River limestone
	Beekmantown limestone

CAMBRIAN

In central Pennsylvania the analogous sequence to the above reads as follows (Butts and Moore 1936) :

SILURIAN	
ORDOVICIAN.....	Juniata formation
	Bald Eagle member at base
	Reedsville shale
	Trenton limestone
	Black River group
	Carlin limestone
	Beekmantown group

CAMBRIAN

* Pulaski age.

† Eden age and older.

Among the formations assigned to the Ordovician system in Pennsylvania, red or other presumably continental units are not common, even if we include the Juniata. Following the usage of Willard and Cleaves (1939), the Juniata is now used to embrace the Bald Eagle conglomerate as its basal member. Upon indirect paleontological evidence this thick unit of central and south-central Pennsylvania has been alternately called late Ordovician and early Silurian. By the theory of periodic sedimentary cycles, it might be grouped with the Ordovician. Yet it was deposited *after* the Taconic Disturbance, which is commonly taken as marking the close of the Ordovician period. The evidence for the origin of these beds according to the above-cited observers is as follows:

"Resumed uplift unaccompanied by folding in the east closed Maysville time. Active, terrestrial erosion recommenced; the succeeding Bald Eagle and Juniata red beds spread westward and northwestward over the shales and sandstones of the Martinsburg and Reedsville. These continental red sediments were syn-genetically colored through derivation from red regolith accumulated on the Martinsburg and older (*cf.* New Jersey) beds during late Maysville time following the Taconic Disturbance. It might be supposed, as was suggested by Willard (1928) that the Tuscarora, Shawangunk, Juniata, and Bald Eagle are in part or wholly the material eroded directly from the east and laid down to the west and northwest immediately following the Taconic Disturbance. Our present understanding of the Bald Eagle, its distribution, and relations to succeeding and underlying beds blasts this theory, since we cannot by such an explanation account for the wide distribution of the Bald Eagle conglomerate across the beveled edges of the Martinsburg shale and sandstones."

The latter portion of the Ordovician of south-central and eastern and central Pennsylvania is made up of dark shales below the Juniata (and Bald Eagle) or the basal Silurian (Tuscarora or Shawangunk). These beds are the Reedsville of our central counties, the Martinsburg of other areas. In the Schuylkill Valley and westward therefrom, the Martinsburg contains a prominent block of red beds relatively high in the sequence. They are Willard's Jonestown beds (1939). A local feature, their age is believed to be pre-Pulaski but post-Eden, that is, they are probably Maysville. Their limited areal extent and short stratigraphic range are partly attributed to their having developed near the positive Harrisburg axis and being today preserved in a long, narrow, east-west trending synclinorium in Berks, Lebanon, and Dauphin counties.

The lithology of the Jonestown beds and their red color point to non-marine conditions of origin, although non-red shales also are present. The red beds are associated also with thin, more or less

lenticular limestones which are barren, platy to submassive and which may alternate with shale interbeds (Miller 1937). They contain cross-bedded, sandy layers, oölite, and an abundance of limestone breccia or edgewise conglomerate. From neither the limestones nor the associated clastic sediments has any fossil, animal, or vegetable been recorded. However, mud cracks, current ripple marks, and probable rain prints occur. The presumptive evidence is for freshwater origin of the limestones and associated red beds. Assuming alternate wet and dry seasons, many of the peculiar features of the limestones might be explained as having originated in playa lakes.

The total maximum thickness of the Jonestown red beds of the Martinsburg is reported (Willard 1939) as approximately 500 feet. It will be observed that they do not mark the close of the Ordovician period but are followed by marine beds of Pulaski age, thus, they far antedate the Taconic Disturbance which came in post-Pulaski time.

The Silurian

Silurian sequences in Pennsylvania are well known and have been expounded fully and clearly by C. K. and F. M. Swartz (1931, 1934, and 1939). The following sequences have been adopted for use here as fairly typical and illustrating differences in sequences between eastern and central Pennsylvania and New Jersey.

Central Pennsylvania	Eastern Pennsylvania	N. Central N. J.
Cayugan series	Cayugan series	Cayugan series
Keyser limestone	Keyser limestone	Decker limestone
Tonoloway limestone	Bossardsville limestone	UNC?
Wills Creek shale	Poxono Island shale (?)	?
Bloomsburg red shale	Bloomsburg red shale	
Niagaran series	Niagaran series	Niagaran series
McKenzie sh. and ls.	Bloomsburg red shale*	Bloomsburg ("Long-
Clinton group	Clinton group†	wood") red shale
Medinan series	Medinan series	Medinan series
Castanea sandstone }	Shawangunk ss. and egl.	Greenpond conglomerate
Tuscarora sandstone		

The occurrence of continental facies in the Silurian is well known and requires no long review nor description here. In north-central New Jersey practically the entire Silurian system is continental, a step farther than any observed conditions in the Pennsylvania sequences. In eastern Pennsylvania the Silurian above the gray, 1800 foot marine

* Contains occasional, thin, green strata.

† Entirely continental facies

Shawangunk conglomerate may be in part sandstone assigned to the Clinton group and of continental origin. These beds are overlain by 1800 to 1900 feet of the Bloomsburg red shale and sandstone, also belonging to the continental facies. Green interbeds occur in the Bloomsburg. Sufficient vegetation was present in this unit to produce thin coaly films (Willard 1938a). Otherwise, organic remains consist of little else than ostracoderms in the Bloomsburg of Perry County (Claypole 1885). The Clinton passes over into marine beds in central Pennsylvania. The Bloomsburg continues the continental facies upward and beyond the limits of the non-marine Clinton. It thins northwestward from the base upward to a mere 30 foot remnant beneath the Wills Creek shale at Mount Union, Huntingdon County (F. M. Swartz 1934). Thus, in eastern Pennsylvania, at Lehigh and Delaware water gaps, the base of the continental facies is late Medinan age (or early Clinton). In central Pennsylvania it has risen to early Cayugan.

In other words, the youngest part of the continental facies of the Silurian, or the highest Bloomsburg red beds, has the maximum areal extent; the oldest beds, early Medinan, the least. As it thins away from the locus of greatest deposition, it gradually displaces successively higher marine beds. The Bloomsburg does not punctuate the close of the Silurian system as now defined in Pennsylvania. It is followed in its thickest and best developed sections by marine shale and limestone of some importance. Of these beds some, the Tonoloway and Keyser formations, are nearly continuous throughout the Upper Silurian exposures of the State. It will be noted that the Keyser has only recently been assigned to the Silurian (F. M. Swartz, in Willard, Swartz, & Cleaves 1939).

The Devonian

The Catskill question is a veteran among the stratigraphic problems of eastern North America. Recently, Willard (1939) reviewed it for Pennsylvania. Like the Bloomsburg facies of the Silurian, the Catskill of the Devonian is a true continental facies which is thickest in the east or southeast and thins toward the west and northwest. The thinning likewise is from the bottom upward. Similarly, too, it displaces successively younger marine units by its progressive off-lap from the region of maximum supply of sediments. It contains approximately 50 percent of non-red beds in the northeastern part of the State, a far larger proportion than did the Silurian. Some hundreds of feet of non-red (green and greenish) strata are interbedded with the

red, not as single, scattered beds, but as thick units. The post-Portage sequence in northeastern Pennsylvania is as follows:

Unit name	Dominant color	Average thickness
Mt. Pleasant	Red	243 feet
Elk Mountain	Green	172 "
Cherry Ridge	Red	485 "
Honesdale	Gray and green	270 "
Damascus	Red	388 "
Shahola	Gray and green	727 "

The Catskill contains very thin coal beds (maximum observed not over two inches). Land plants, however, are quite abundant locally, as scattered stems and leaves or as mats of vegetation. The fresh-water mollusc, *Amnigenia*, or other forms of similar habits, is rare. Ostracoderms are present but not generally common in Pennsylvania. Unlike the Bloomsburg facies which fails before the Silurian-Devonian boundary is attained, the Catskill continues through to the very end and is succeeded in most exposures in Pennsylvania by non-marine Mississippian beds. Throughout, the continental sequence may be observed to pass over into marine beds, and both the red and gray elements are transitional with salt-water-formed strata. The change and its progressive or transitional character may be tabulated (marine beds cited by groups only):

N. W. Penna.	Central Penna.	Eastern Penna.	N. Central N. J.
Connewango	Catskill	Catskill	
Conneaut	Catskill	Catskill	Eroded
Canadaway	Catskill	Catskill	
Chemung	Chemung	Catskill	
	Portage	Portage	Catskill
Concealed	Hamilton	Hamilton	Hamilton
	Onondaga	Onondaga	Onondaga
	Oriskany	Oriskany	Absent
	Helderberg	Helderberg	

The Mississippian

The following sequences serve to illustrate the Mississippian successions in Pennsylvania. The thicknesses are disregarded, although there is much disparity among the several units.

S. W. Penna.	Allegheny Front	Susquehanna Valley	Pottsville
Mauch Chunk*	Mauch Chunk	Mauch Chunk	Mauch Chunk
Greenbriar ls	Mauch Chunk	Mauch Chunk	Mauch Chunk
Loyalhanna "ls"	Loyalhanna "ls"	Mauch Chunk	Mauch Chunk
Pocono†	Pocono	Pocono	Pocono
Pocono	Riddlesburg sh	Pocono	Pocono
Pocono	Pocono	Pocono	Pocono

All of our eastern strata which are assigned to the Mississippian system are non-marine and are divided between the gray Pocono below with an average thickness of 1500 feet, and the red Mauch Chunk above, which has a maximum thickness in the Lehigh Valley and near Pottsville of about 3000 feet but is usually thinner. The eastern Pocono carries few or no red beds; the Mauch Chunk is practically nothing but red sandstone and shale throughout. However, from the region of the Allegheny Front and Broad Top Mountain in Huntingdon County westward, marine Mississippian beds begin to be intercalated into the sections. The marine Riddlesburg shale splits the Pocono near its middle. At or near the base of the Mauch Chunk, the Greenbriar limestone and the Loyalhanna "limestone" appear. The latter is more often a calcareous sandstone. The marine Greenbriar is usually highly fossiliferous. The Loyalhanna is commonly reported barren, but recently, marine Mississippian foraminifera and ostracoda have been collected from it. The thin, marine, Lower Mississippian sequence of northwestern Pennsylvania is taken to be the partial correlate of the lower Pocono.

The Pocono is very like the non-red sandstones of the preceding Catskill, particularly the Honesdale with which it was confused for years. The Mauch Chunk is a near lithologic replica of such Devonian red units as the Damascus and Cherry Ridge. From such superficial resemblance, it would appear that the Pocono-Mauch Chunk sequence is merely a continuation of Catskill conditions in a more aggravated form.

Fossils of the continental Pocono are confined to plants which are often abundant enough to produce poor coal seams, ranging from a mere film up to beds four feet thick. The Mauch Chunk is particularly barren save for rare plant fossils and the foot tracks of amphibia and perhaps reptiles which have been discovered locally in some abundance.

* Mauch Chunk here applies to the red continental beds of the Mississippian.

† Pocono here applies to the non-red (gray) continental beds of the Mississippian.

The salt-water-freshwater relations of our Mississippian formations are less apparent than were those of the Devonian or Silurian. That an incomplete transition west and northwest took place in Pocono time is quite certain, but the general concealment of the system beneath our bituminous coal fields hides nearly all details. The Mauch Chunk is different from older continental beds for it shows peculiarities not earlier recognized. It appears probable from thickness data that there were at least two loci of maximum deposition during Mauch Chunk time. One of these lay approximately in the region of the southern Anthracite Field and the other in south-central Pennsylvania and west of or near the Allegheny Front. From these, the beds thin away fanwise. The two great fans coalesce in the Susquehanna Valley. The Mauch Chunk displays yet another unique feature in its behavior northwestward from the loci of deposition. Eastward from the Anthracite Fields it has been eroded, but to the north instead of passing over into marine beds it thins out to a feather edge. Therefore, where it has disappeared, the Pocono may be followed directly by the Pottsville of the Pennsylvanian system. From its western maximum locus, the Mauch Chunk likewise fans out to the north but does show a tendency to become marine westward.

The Pennsylvanian and Permian

Only a word should be said here of the Pennsylvanian and Permian systems. No Permian is present except in our southwesternmost counties. Throughout the Anthracite area of the east, non-red continental beds are the rule for the Pennsylvanian. Few marine fossils have been discovered; red beds are practically absent. The vegetation was prolific. The thick coal measures witness its abundance. In general the Pennsylvanian suggests recurrent but accentuated Pocono conditions. However, there is no succeeding red facies to correspond to the Mauch Chunk. West of the Allegheny Front the Pocono-like character of the Pennsylvanian continues. Nevertheless, like the western Pocono, one finds intercalations of an occasional marine bed, mostly limestones, among the soft-coal measures. There was a very minor freshwater to salt-water shift westward.

The Triassic

Though beyond the Paleozoic limit, it is nevertheless interesting to observe the Triassic (Newark group) as an *apparent* reversion to the red type of continental sedimentation in eastern Pennsylvania. In its vertebrate tracks, its plants, its dominance of red sandstone and shale, the Triassic is peculiarly reminiscent of the Mauch Chunk. In history

and origin, however, it bears little analogy to the older red beds of any Paleozoic system.

THE FACIES SHIFTED

Evidently, from Ordovician times on throughout the Paleozoic, there was a more or less regularly recurrent tendency in Pennsylvania for the sedimentary facies to shift from the marine to continental phase of deposition. The shifts progress from southeast, northwestward generally. As the continental facies advance and displace the marine, there is a concomitant thinning of the non-marine units northwestward, that is, in the direction of off-lap or away from the source of supply. Within the marine sequences themselves, there are other unmistakable shifts of major significance, but these are beyond the scope of this treatise. We shall consider only the marine-to-continental shifts in chronological order.

It seems wise to regard the Juniata red beds as a special instance which hardly fits into the scheme. The Juniata (including the Bald Eagle) thins from central Pennsylvania south and east. Its greatest development, most typical form, and largest amount of coarsest conglomerate (the type Bald Eagle) are found in our central counties. It shows no lateral transition into marine beds and may be separated from the underlying marine Martinsburg or Reedsville shales of Eden age by an unconformity or disconformity (Taconic). The Juniata may or may not intergrade with the overlying Shawangunk or Tuscarora gray or white sandstone and conglomerate. For these reasons, the deduction is that at the close of the Ordovician, the Juniata was deposited more or less locally as material eroded from northeastern or even northern areas. It appears to bear out the stratigraphic maxim that continental deposits thicken toward the source of supply.

The Jonestown red beds are an entirely different story from that of the Juniata. They conformably overlie dark marine shales of known (Eden) age and are succeeded by beds which carry Pulaski fossils. Laterally, their relations are surmised rather than observed, for they are confined to an east-west synclinorium reaching from the Susquehanna at least to the Schuylkill. To the north of the synclinorium they have been eroded and are not known to reappear in central Pennsylvania. Likewise, they are believed to have been worn away to the south. Their relation to the Slate Belt sequence of eastern Pennsylvania is not as yet fully determined.

From all observations and from analogy with later sequences, only the red Jonestown beds are the true representation of the Ordovician

continental facies in *sensu stricto* in Pennsylvania. Because such a small remnant is preserved, and since the distribution northwestward appears to have been very limited, it is deduced that the source of these Ordovician red beds was at a rather remote distance to the southeast. The fact that more marine beds of Ordovician age follow these red ones is no deterrent from the premise that they are the chief expression of the continental Ordovician. This situation recurs in the Silurian system, as we shall presently describe.

In north-central New Jersey the Green Pond Mountain area embraces a long, narrow band of Paleozoic formations preserved by down-folding and down-faulting (Kümmel & Weller 1902). It includes practically the entire Silurian system which is here chiefly of freshwater origin. In northeastern Pennsylvania, typically at the Delaware Water Gap section, the upper half of the system is built up of red shale plus a few, scattered greenish strata (Swartz & Swartz 1931). This is the Bloomsburg continental facies. The facies actually transcends the base of the true Bloomsburg into the upper part of beds which have been assigned to the Shawangunk, but may be truly of Clinton age and of continental origin. The Silurian continental facies proceeds to thin northwestward and westward from the Delaware Valley. Like the Ordovician continental facies, the Silurian beds fail to continue to the close of the period. Unlike the Ordovician, the lateral transition is so well-marked that practically the entire Silurian system is observed to pass over, even before our eyes, from the freshwater to the salt-water type of sediments.

If we do with the Devonian as we did with the Silurian and widen our observations to include north-central New Jersey, swing thence westward, and trace the changes in the sequences across Pennsylvania and perhaps go through even as far as the northwestern corner of the State, the picture yields a nearly complete transition between the two facies, the freshwater on the east, the marine to the west. In the farthest east section in northern New Jersey, the Oriskany and Helderberg groups are absent, but marine beds of Onondaga and Marcellus ages are recognized. Above, the remainder is all continental, Catskill facies. Thus, the Lower Devonian fails to attain quite the lower limit shown by the Silurian analogue. The latter attains more nearly to perfection of change.

On the other hand, the top of the highest Devonian is fully occupied by the continental facies in northeastern Pennsylvania, thus going beyond the Silurian in the completeness of its uppermost displacement. Throughout its succession, the Catskill continental facies displaces

successively from southeast to northwest the Hamilton, Portage, Chemung, Canadaway, Conneaut, and Conewango groups of marine Devonian formations.

Throughout the Catskill sequence, the most astonishing feature and that which at once sets it off from the older continental facies, but ties it in with later ones, is the large proportion of non-red elements. This situation cannot be too fully emphasized because its interpretation bears directly upon the origin of these beds.

The Mississippian transition is almost absent from Pennsylvania. By far the majority of exposed beds of this system over wide areas is non-marine. Nevertheless, the change begins to show itself in south-central Pennsylvania. Very little of the transition has actually taken place where the system plunges westward under the Allegheny Front. Where the Mississippian again rises to the surface as along the Chestnut Ridge anticline in Fayette County, the transition has progressed a pace but the displacement is still a minor factor.

Mention was just made of the important change shown by the Devonian continental facies in the introduction of about 50 percent of non-red material. A proportionally similar admixture is found in the Mississippian in the percentage relations between gray Pocono and red Mauch Chunk. In the east, the red is about twice as thick as the gray, but the former thins so rapidly from supply source that this proportion is soon reversed. In general, and on the average, the proportions appear to be about half-and-half red to gray. A final step in this remarkable shift within the continental sediments is its culmination in the Pennsylvanian and Permian systems where red beds are all but unknown.

Summarizing these changes, we have observed, first, the progress from all red continental beds in the Ordovician to essentially completely non-red continental beds in the Pennsylvanian and Permian:

TABLE 1
CONTINENTAL COLOR RELATIONS

System	Percentage of red beds	Percentage of non-red (gray or green) beds
Permian and Pennsylvanian	1	99
Mississippian	50	50
Devonian	50	50
Silurian	90	10
Ordovician	99	1

Second, the continental to marine transition may be tabulated:

TABLE 2
DEGREES OF CONTINENTAL TO MARINE TRANSITION

System	Completeness of transition
Permian	none
Pennsylvanian	trace
Mississippian	little
Devonian	complete
Silurian	complete
Ordovician	record incomplete

The organic remains of the continental units are suggestive. The plant evolution is marked by increase in abundance of land floras. The animals illustrate the gradual conquest of the terrestrial environment by the vertebrates.

TABLE 3
FOSSILS OF THE CONTINENTAL UNITS

System	Organic remains
Permian and Pennsylvanian	skeletal remains of land vertebrates, insects, abundant coal-forming vegetation
Mississippian	tracks of land vertebrates, coal-forming vegetation
Devonian	ostracoderms, freshwater molluscs, land plants
Silurian	ostracoderms, freshwater (land?) plants
Ordovician	no fossils recorded

The source of the continental sediments is assumed always to have been Appalachia (with, to be sure, the possible exception of the Juniata which we ruled out as non-orthodox). The ancient land mass supplied the detritus, and it appears that the actual direction from which these clastics came varied little from period to period. Roughly, this may be tabulated as follows:

TABLE 4
DIRECTION OF SOURCE OF SUPPLY OF SEDIMENTS OF CONTINENTAL FACIES

System	Direction
Permian and Pennsylvanian	east and southeast
Mississippian	southeast
Devonian	southeast and east
Silurian	southeast
Ordovician	southeast and south (?)

Emphasis has been laid upon the color shift among the continental facies. There is little of this alternation of color in the Ordovician. It is suggested in the Silurian, in the Devonian at least three major repetitions of the process are recorded, and the Mississippian system gives us one complete swing from non-red to red followed by a return oscillation in the vast, thick, non-red Pennsylvanian and Permian systems.

To produce such an alternation of conditions, we presume alternation of intervals of uplift and erosion of Appalachia. What the nature of this uplift was, we have only vague hints. It may have been largely epeirogenic, yet in the case of the Taconic Disturbance it was orogenic. The alternation of colors, which began in the Silurian and continued on an even grander scale through Mississippian-Pennsylvanian-Permian sequences, points to a rising crescendo of movements in Appalachia. Perhaps the Devonian beds register indirectly the effects of the Acadian disturbance. Certainly, the movements hinted in the Silurian, clearly developed in the Middle Devonian and increasing in intensity from then to the close of the Paleozoic mark the culmination of convulsions in Appalachia which we appropriately call the Appalachian Revolution.

The Geosyncline Migrated

The sequence of continental facies particularly in eastern Pennsylvania appears to illustrate a steady increase over the marine of the development of this phase of sedimentation from Ordovician through the close of the Paleozoic era. Is this more apparent than actual? Is it not more probable that the continental facies may at some locality or another have had a nearly equal development in all systems; that we would find this true were they completely preserved for us? The Silurian and Devonian illustrate this when we include in our observations the remnants of these systems in north-central New Jersey. By analogy, it is highly probable that the Ordovician would have shown a like condition were its more easterly portions preserved. The apparent increase in intensity or degree of continental deposition from older to younger systems in eastern Pennsylvania may be attributed to a westward or northwestward migration of the Appalachian geosyncline. In the Ordovician its eastern shore lay somewhere east of us, we know not where. In Silurian and Devonian times it crossed northern New Jersey and probably southeastern Pennsylvania. By Mississippian time it had passed westward and was entirely outside of Pennsylvania during part of the period. In the still later periods, marine sedimentation was no longer a factor within the State. Such

a migration coincides with our postulated, progressively accentuated uplift with possible folding in Appalachia. Add to this, that uplift was so pronounced and the supply for sedimentation so abundant toward the close of the Paleozoic that isostatic adjustment beneath the geosyncline may not have then kept pace with the increased piling up of sediments, and we have an added factor in the westward migration of the area of sedimentation. Only in the final collapse of Appalachia and the production of the down-faulted areas or graben of Triassic time did Paleozoic conditions of sedimentation end.

The presence of red beds is doubtless in part an expression of the climate. Proof is negative for any vast climatic change during the Paleozoic in-so-far as our sediments are concerned. That it may have been cool at times is indicated by evidence for the presence of ice in Middle Devonian seas (Willard 1939). In contradistinction to this is the occurrence in marine beds throughout the sequence from the Ordovician into the Pennsylvanian of corals. The gradual development of coal-forming vegetation is more truly an expression of the process of organic evolution than of climatic change. There is no obvious reason for lack of coal-forming floras in pre-Pennsylvanian times other than the very simplest explanation of all; in those earlier periods no suitable plants capable of coal-producing in abundance had evolved. Ostracoderms are present in the Silurian and likewise in the Devonian red beds. Their habits are assumed to have been similar. Four-footed animals crawled the land in Mississippian as well as Pennsylvanian and Permian times, have been dubiously identified from the Devonian. Again, their habits must have been similar. Rain prints and mud cracks generally present in our red beds add, to the organic data, paleoclimatic testimony. Appalachia's relief must have been reasonably high intermittently, but at other times it probably stood low and was partially sea-flooded. There is no reason to suppose that the prevailing westerlies were not functional, therefore, moisture may well have been plenty on the west coast of Appalachia during most of the Paleozoic. The comparatively warm, moist climate not only affected the life of the sea and the land. It induced rapid accumulation through chemical weathering of residual soils which by analogy with recent conditions should have been reddish.

THE SHIFT OF THE FACIES SHIFTS

Throughout the Paleozoic there have been in Pennsylvania recurrences of continental facies tending repeatedly to displace their marine contemporaries. The recurrence may be more or less cyclic. These displacements have been referred to as continental to marine facies shifts. Actually, the freshwater beds themselves display two marked cumulative changes independent of their displacement of the marine beds. First, there is the progressive, chronological evolution from red to non-red rock color (TABLE 1). Second, there is a progressive shift in position or westward and northwestward spreading at each recurrence of continental beds (FIGURE 2). Each successive continental facies after the Jonestown,—the Bloomsburg and Clinton, the Catskill, the Mauch Chunk and Pocono, the Pennsylvania and Permian,—spreads farther west or northwest than did its predecessor. In each there is a progressively greater encroachment of freshwater strata upon the Appalachian geosynclinal seaway. The continental-marine boundaries are themselves an expression in their recurrence of a still wider shift, an era-long migration, a change of momentous proportions. Since all of the sediments, save for some reworking of older beds, were presumably derived directly from Appalachia, is not this shift of shift an expression of vast, slow changes affecting that ancient land-mass during post-Cambrian time?

A key to this lies in the alternating red with non-red units and the percentage of change between them. If, as postulated, the red is reworked regolith, but the non-red is derived from more deeply eroded, fresher rock, then must there have been a succession of uplifts separated by pauses affecting the relief of Appalachia. During each pause, weathering produced red residuals. Each uplift rejuvenated streams. Accelerated erosion ruled. Red beds were formed in piedmont and coastal plain. They intergraded beyond with non-red marine strata. Further erosion cut into fresh bed rock. Non-red continental strata were deposited over the preceding red ones. They, too, graded laterally into marine beds. The process was repeated at least five times, each recurrence "nobler than the last". With each repetition the red element tended to grow less conspicuous, thinned in proportion to the non-red, the green or more commonly gray elements. Conversely, the gray portions waxed thicker and dominated until by Pennsylvanian-Permian time all red tended to vanish from the sections.

Historical geologists speak of Taconic and Acadian Disturbances, of the Appalachian Revolution. Are these intervals of orogeny dis-

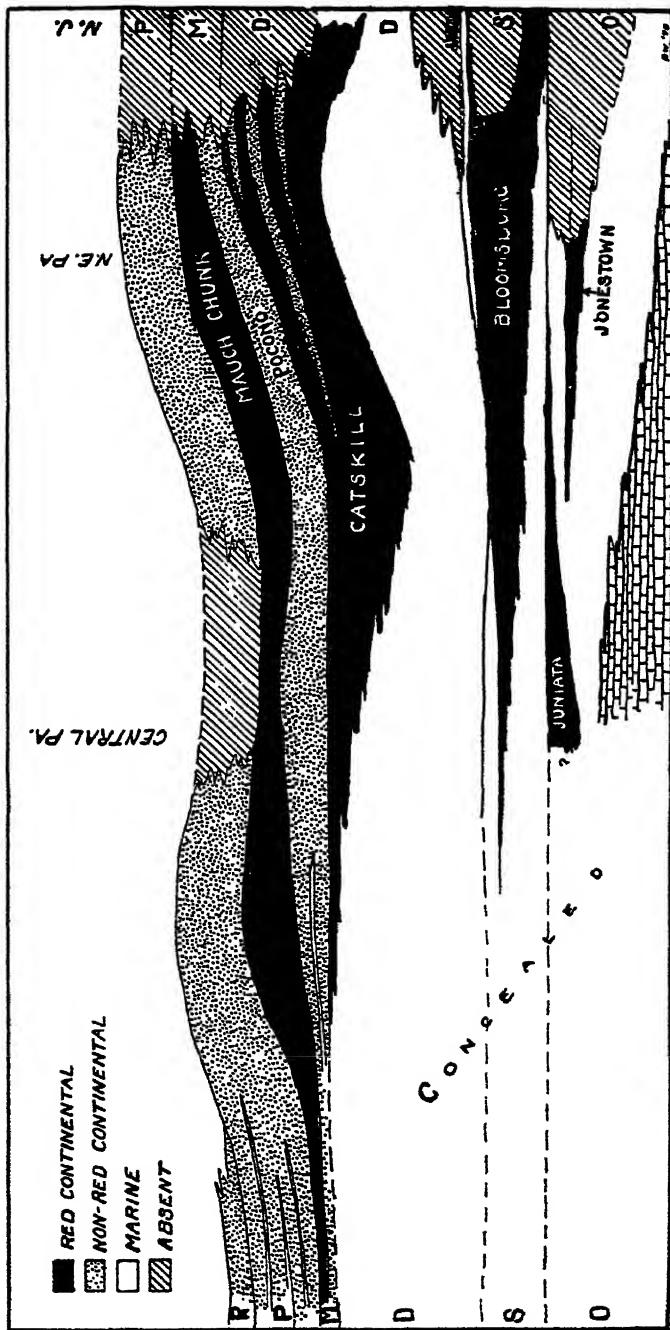


FIGURE 2. Diagrammatic cross section to illustrate the shift of the facies shifts during the Paleozoic in Pennsylvania. Thicknesses are approximate, the maximum for the entire sequence being nearly 28,000 feet. Horizontal distance approximately 375 miles.

tinct entities? In these continental facies, whose character we have been at some length to expound and explain, have we a reflection of orogeny or epeirogenic movements in Appalachia? If our interpretation of the sedimentary record is correct, there were a few distinct intervals of diastrophism marked off by terms of quiescence. There was progression toward a culmination in epeirogenic and orogenic activity whose acceleration the sediments reflect. Its final, grand expression is truly our Appalachian Revolution. On these grounds, the Appalachian Revolution began during the Ordovician period, at least as far back as Eden or Maysville time.

BIBLIOGRAPHY

- Butts, Charles; & Moore, E. S.
1936. Geology and mineral resources of the Bellefonte quadrangle, Pennsylvania. U. S. Geol. Surv. Bull. 855: 1-112.
- Claypole, E. W.
1885. A preliminary report on the palaeontology of Perry County. Penn. 2nd Geol. Surv. Rept. F2: 1-437.
- Kümmel, Henry B.; & Weller, Stuart
1902. The rocks of the Green Pond Mountain region. N. J. Geol. Surv. Ann. Rept. 1901. 1-46.
- Miller, R. L.
1937. Martinsburg limestones in Pennsylvania. Bull. Geol. Soc. Am. 48: 93-112.
- Swartz, C. K.; & Swartz, F. M.
1931. Early Silurian formations of southeastern Pennsylvania. Bull. Geol. Soc. Am. 42: 621-662.
- Swartz, F. M.
1934. Silurian sections near Mount Union, central Pennsylvania. Bull. Geol. Soc. Am. 45: 81-134.
1939. Keyser limestone and Helderberg group in the Devonian of Pennsylvania. In, "The Devonian of Pennsylvania," by Bradford Willard, F. M. Swartz, & A. B. Cleaves. Penn. Topog. Geol. Surv. Bull. G19: 29-91.
- Willard, Bradford
1938a. Evidence of Silurian land plants in Pennsylvania. Proc. Penn. Acad. Sci., 12: 121-124.
1938b. A Paleozoic section at Delaware Water Gap. Penn. Topog. Geol. Surv. Bull. G11: 1-35.
1939. Ordovician shales of southeastern Pennsylvania. Proc. Penn. Acad. Sci. 13: 126-133.

_____; & Cleaves, A. B.

1938. A Paleozoic section in southcentral Pennsylvania. Penn. Topog. Geol. Surv. Bull. G8.
1939. Ordovician-Silurian relations in Pennsylvania. Bull. Geol. Soc. Am. 50: 1165-1198.

_____; Swartz, F. M.; & Cleaves, A. B.

1939. The Devonian of Pennsylvania. Penn. Topog. Geol. Surv. Bull. G19.

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

VOLUME XL, ART. 5. PAGES 289-482

DECEMBER 31, 1940

DIELECTRICS*

BY

WILLIAM O. BAKER, J. D. FERRY, RAYMOND M. FUOSS, PAUL M. GROSS,
MARCUS E. HOBBS, JOHN G. KIRKWOOD, S. O. MORGAN,
HANS MUELLER, J. L. ONCLEY, HERBERT A. POHL,
J. SHACK, CHARLES P. SMYTH, J. H. VAN VLECK

CONTENTS

	PAGE
INTRODUCTION. BY CHARLES P. SMYTH.....	291
THE INFLUENCE OF DIPOLE-DIPOLE COUPLING ON THE DIELECTRIC CONSTANTS OF LIQUIDS AND SOLIDS. BY J. H. VAN VLECK.....	293
THE LOCAL FIELD IN DIELECTRICS. BY JOHN G. KIRKWOOD.....	315
THE DIELECTRIC ANOMALIES OF ROCHELLE SALT. BY HANS MUELLER.....	321
ROTATION OF SOME LARGE ORGANIC MOLECULES. BY S. O. MORGAN.....	357
THE DIELECTRIC PROPERTIES OF PROTEIN SOLUTIONS. BY J. L. ONCLEY, J. D. FERRY, AND J. SHACK.....	371
POLARIZATION MEASUREMENTS ON CARBOXYLIC ACIDS IN DILUTE SOLUTION IN NON-POLAR SOLVENTS. BY HERBERT A. POHL, MARCUS E. HOBBS, AND PAUL M. GROSS.....	389
THE ELECTRICAL PROPERTIES OF POLYVINYL CHLORIDE PLASTICS. BY RAYMOND M. FUOSS.....	429
THE DIELECTRIC CONSTANTS OF SOME ORGANIC CRYSTALS AND GLASSES. BY WILLIAM O. BAKER AND CHARLES P. SMYTH.....	447

* This series of papers is the result of a conference on Dielectrics held by the Section of Physics and Chemistry of the New York Academy of Sciences, April 14-15, 1939.
Publication made possible through a grant from the income of the Permanent Fund and the Publication Fund.

COPYRIGHT 1940
BY
THE NEW YORK ACADEMY OF SCIENCES

INTRODUCTION

BY CHARLES P. SMYTH

From the Department of Chemistry, Princeton University, Princeton, New Jersey

In 1912 Debye explained dielectric constants by the theory that the molecules of some substances contained permanent electric dipoles, which caused the molecules to orient in an electric field. Some criticism was directed against this theory and a little experimental work was done to obtain support of it, but, for the most part, the theory lay dormant for ten years. Then, use was made of it for the calculation of a few molecular dipole moments, which were found to be consistent with what might be expected from the probable structures of the molecules. The experimental foundation of the theory was soon firmly established and the subsequent quantum mechanical treatment led to a relation between dielectric constant and dipole moment identical for practical purposes with that which Debye had derived on the basis of classical mechanics. The initial successes of the attempts to correlate dipole moments and molecular structures were preliminary to a rapid extension of the use of the dipole moment as a tool to investigate molecular structures and, in the ensuing years, several hundred dipole moment values were obtained from dielectric constant measurements and interpreted as evidence concerning molecular structures.

Because of the effects of the dipoles upon their neighbors, it was found that the dielectric constant was dependent upon molecular interaction and could, therefore, be used as a means of investigating this interaction. As the dielectric constant of a substance with polar molecules depends upon the freedom of these molecules to orient in an externally applied field, measurements of dielectric constant have recently been used to investigate molecular freedom in crystals and glasses. When the time required for the molecules to orient in an externally applied field is of the magnitude of that required for the establishment of the field, the dielectric constant depends upon the frequency of the alternating current with which it is measured and an energy absorption accompanies the flow of current. This so-called anomalous dispersion may have an important effect upon the properties of a commercial dielectric. As the "relaxation time" or time required for the molecules to revert practically to a random distribution after removal of the applied electric field depends upon the

size of the molecule, the measurement of the anomalous dispersion of the dielectric constant provides a means of determining the approximate sizes of large molecules. Studies of the effects of dipoles upon dielectric behavior may thus yield information of importance to the electrical engineer and lead the biochemist to an increased knowledge of the sizes and shapes of large molecules. A wide variety of not always closely related phenomena may thus be examined in connection with the subject of "Dielectrics."

THE INFLUENCE OF DIPOLE-DIPOLE COUPLING ON THE DIELECTRIC CONSTANTS OF LIQUIDS AND SOLIDS

BY J. H. VAN VLECK

From Harvard University, Cambridge, Massachusetts

Until very recently, the dielectric behavior of liquids and solids has been more or less of an enigma, but in the last two or three years, considerable progress has been made towards a better understanding of this interesting subject.

THE LORENTZ LOCAL FIELD AND ITS CONSEQUENCES

The starting point of all discussions on dielectric constants is always the formula of Lorentz¹ for the local field, *viz.*,

$$E_{local} = E + \frac{4\pi}{3} P. \quad (1)$$

The second term in (1) is an expression of the fact that the coupling between the molecular dipoles, which is particularly important at high concentrations, tends to make them parallel. Thus the dipole-dipole interaction "amplifies" the influence of the applied field E , and makes the effective field E_{local} acting upon the molecule larger than E . The polarization P per cm³ is equal to $N\gamma E_{local}$, where N is the number of atoms per cm³, and γ is the specific polarizability of a single atom or molecule. The standard theory of Langevin and Debye,² gives

$$\gamma = \alpha + \frac{\mu^2}{3kT}. \quad (2)$$

Here the first term is independent of temperature, and represents the contribution of the induced polarization. The second is due to the alignment of permanent dipoles by the applied field, and is found only in polar molecules, whose dipole moment is denoted by μ . Eq. (2) was originally derived in classical theory. It lost its validity in the old (1913) quantum theory, but detailed analysis³ shows that it is restored in the new or true (1926) quantum mechanics provided the

¹ Cf., for instance, Lorentz, H. A. "The Theory of Electrons." Note 54.

² Langevin, P. *Jour. de Physique*. 4: 678. 1905. Debye, P. *Physikal. Zeit.* 13: 97. 1912

³ Van Vleck, J. H. "The Theory of Electric and Magnetic Susceptibilities." Chap. 7.

spacing of the energy levels (other than vibrational) be small or large compared with kT , a proviso usually satisfied in the study of dielectric constants though often not in the theory of magnetism. From (1) and (2) one derives in a well-known fashion the Clausius-Mossotti relation

$$\frac{M \epsilon - 1}{\rho \epsilon + 2} = \frac{L}{3} \left(\alpha + \frac{\mu^2}{3kT} \right), \quad (3)$$

where ρ is the density, M is the molecular weight, L the Avogadro number, and ϵ is the dielectric constant, given by

$$\epsilon - 1 = 4\pi P/E = 4\pi N \gamma E_{local}/E.$$

Since the right side of (3) is independent of ρ , the characteristic ratio on the left side of (3) should, at given temperature, be independent of the pressure and density.

It is usually stated that (3) can be applied successfully to gases, whether polar or non-polar, to non-polar liquids and solids, and to dilute solutions of polar materials in non-polar solvents. By and large this assertion is true, though certain exceptions usually come to light in refined theory or experiment. One of them is the translational fluctuation effect in gases, which is treated by Kirkwood⁴ in the following paper, and which owes its origin to the fact that in gases the molecules are not evenly spaced, as presupposed in deriving (3). Internal free rotation of the molecule, of which classic example is $C_2H_4Cl_2$, can sometimes give deviations from (3). It is hard to understand theoretically why (3) applies quite as well as it does empirically to non-polar liquids and solids, inasmuch as a detailed quantum-mechanical investigation shows that at high concentrations (3) is rigorous, even for non-polar media, only if one uses the harmonic oscillator model,⁵ which cannot be regarded as a true picture of a real Rutherford atom. Experimentally, in dilute solutions, the contribution of the polar solute appears to be slightly influenced by the nature of the non-polar solvents. These corrections are, of course, all effects of a comparatively high order.

It has, on the other hand, long been recognized that (3) fails completely in pure polar liquids or solids. The difficulties can be seen in a particularly succinct form if one solves the equations for the dielectric constant itself rather than the Clausius-Mossotti ratio. For simplicity, let us neglect the induced polarization α . We then have

⁴ Kirkwood, J. G. Jour. Chem. Phys. 4: 592. 1936.

⁵ Van Vleck, J. H. Jour. Chem. Phys. 5: 556. 1937.

$$\frac{\epsilon - 1}{4\pi} = \frac{N\mu^2}{3k(T - T_c)}, \quad (4)$$

with $T_c = \frac{4\pi N\mu^2}{9k}. \quad (5)$

According to (4), there is a critical or "Curie" temperature T_c at which ϵ becomes infinite. This does not mean that the dielectric constant really increases without limit. Instead at temperatures inferior to T_c , it is not allowable to treat the polarization effects as linear in the field strength, as presupposed in obtaining (3) or (4). Below T_c , saturation effects should enter, and one should expect the electrical analogue of ferromagnetism, i. e. hysteresis, remanence, hysteresis, etc. Actually, such a "ferro-electric" behavior is practically unknown, being confined to Rochelle salt, and one or two other highly anisotropic crystals.^{5a} Even these materials show this deportment only for certain particular directions. Such exceptions are not unreasonable theoretically, for in anisotropic substances, only the average of the Lorentz factor over all directions need be $4\pi/3$, and in a particular direction the factor may be so large as to make the ferromagnetic tendency of the dipole-dipole coupling abnormally important. In the present paper we shall consider only isotropic media, such as liquids, where the anomalous behavior predicted by (4) does not arise. It is convenient to call the vanishing of the denominator of (4) a $4\pi/3$ catastrophe, for the difficulty comes from the presence of the factor $4\pi/3$ in (1). Without the second term of (1), the denominator of (4) would be simply $3kT$, and there would be no saturation difficulties. If the critical Curie temperatures T_c given by (5) were unobtainably low, the difficulty of the $4\pi/3$ catastrophe would not be serious. Actually one calculates according to (5) that T_c equals 1200° and 260° for H_2O and HCl respectively, whereas it is a matter of common knowledge that water and hydrochloric acid never behave ferro-electrically.

THE EMPIRICAL FORMULAS OF WYMAN AND OF VAN ARKEL AND SNOEK

Since the $4\pi/3$ catastrophe is not actually found, one seeks to obtain both empirical and theoretical formulas which will avoid its

^{5a} Another anisotropic material which behaves ferro-electrically is potassium dihydrogen phosphate, and an interesting theory of the dielectric properties of this substance has just been developed by Slater, J. C. *Jour. Chem. Phys.* (in press).

occurrence. First we will mention some empirical expressions. One of these is due to Wyman,⁶ and has the form

$$\frac{\epsilon - 1}{8.5} = \frac{4\pi N}{3} \left(\alpha + \frac{\mu^2}{3kT} \right). \quad (6)$$

Wyman has shown that (6) accounts for the orders of magnitude of the dielectric constants of many substances at room temperatures. Another, rather more complete, empirical formula is that of two Dutchmen, van Arkel and Snoek,⁷ *viz.*:

$$\frac{M \epsilon - 1}{\rho \epsilon + 2} = \frac{4\pi L}{3} \left[\alpha + \frac{\mu^2}{3kT + (4\pi C/3) N \mu^2} \right]. \quad (7)$$

Here C is an empirical constant which, however, usually turns out to lie between 1 and 1.7. If $(4\pi Na + 3C^{-1}) < 3$ the $4\pi/3$ catastrophe is avoided, for then the right side of (7) never reaches as high a value as that M/ρ corresponding to $\epsilon = \infty$. Van Arkel and Snoek show that with fixed C , formula (7) represents surprisingly well the dependence on temperature, and particularly on concentration for solutions of polar materials in non-polar solvents. As an example, some figures are given in TABLE 1, taken from Müller,⁸ for various concentrations of nitrobenzene in CCl_4 .

TABLE I
MOLAR POLARIZATION OF NITROBENZENE IN CARBON TETRACHLORIDE SOLUTION

$N \times 10^{21}$	0	.132	.597	1.31	3.60	5.90
P^{OR} (obs)	369	338	274	212	127	95.5
P^{OR} (calc)	369	340	267	204	125	95.5

Here N is the number of nitrobenzene molecules/cc. The value 5.90×10^{21} relates to pure nitrobenzene. The expression P^{OR} is the part of the "molar polarization" due to orientation of the permanent dipoles, defined by

$$P^{OR} = \frac{\epsilon - 1}{\epsilon + 2} \frac{M}{\rho} - \frac{4\pi L \alpha}{3}. \quad (8)$$

The calculated values of P^{OR} are on the basis of $C = 1.32$. We must, however, mention that with increasing experimental refinement, much of the apparent close applicability of (7) is often lost, particularly at

⁶ Wyman, J. Jour. Am. Chem. Soc. 58: 1482. 1936.

⁷ Van Arkel, A. E., & Snoek, J. L. Trans. Faraday Soc. 30: 707. 1934.

⁸ Müller, F. H. Physikal. Zeit. 38: 498. 1937.

low concentrations. For instance, Smyth⁹ reports that for ethyl bromide in hexane at 30° the quantity C , instead of remaining constant, really varies from 1.3 to 5.0. The high value applies to a concentration .03, and C does not exceed 1.4 until the concentration is less than .50. Smyth also finds some temperature variation of C . For instance, in pure ethyl bromide, C equals 1.33 at 30° C and 1.49 at -90° C. On the whole, though, (7) is a remarkably good first approximation.

FOWLER'S AND DEBYE'S THEORY OF HINDERED ROTATION

So much for empirical formulas. It is obviously more satisfactory to find a theoretical basis for avoiding the $4\pi/3$ catastrophe. One very celebrated attempt of this sort is the hypothesis of hindered rotation, proposed in this connection independently by Fowler¹⁰ and by Debye.¹¹ The basic idea of their theory is that in the solid or liquid states the molecules are not free to turn, but instead are resisted by an internal field which is superposed on (1) and which is supposed to arise from interatomic forces. It is essential that this internal field be unilateral, rather than bilateral, *i.e.* have a potential function whose period is 360° rather than 180° when the molecule is rotated, for a simple calculation,¹² too often overlooked, shows that internal fields with bilateral symmetry have absolutely no effect in reducing the dielectric constant. The internal fields are supposed directed at random, as otherwise the specimen would have a residual polarization even in the absence of an applied field. If the internal field is sufficiently large, the difficulty of a $4\pi/3$ catastrophe is avoided, and the dielectric constant is reduced to its empirical value.

The hypothesis of hindered rotation has many arguments in its favor. The necessary internal potentials usually correspond to an energy of the order $10kT$ for "turning over" the molecule, which is reasonable. The disappearance of the "ferro-electric" behavior of Rochelle salt below a certain critical temperature receives a natural explanation¹⁰ in terms of the sudden onset of hindered rotation. (This use of hindered rotation may be correct, even though we shall later see that its application to ordinary isotropic substances is questionable.) Müller⁸ has shown that if the internal field is very large, the

⁹ Smyth, C. P. Jour. Phys. Chem. 43: 131. 1939.

¹⁰ Fowler, R. H. Proc. Roy. Soc. 149A: 1. 1935.

¹¹ Debye, P. Physikal. Zeit. 36: 100, 193. 1935.

¹² Cf., for instance, Mueller, H. Phys. Rev. 50: 547. 1936.

model with hindered rotation yields the empirical formula (7) of van Arkel and Snoek, though only incompletely, *viz.* with only the second member of the denominator $3kT + (4\pi C/3)N\mu^2$. A further success of the theory is found in Debye's calculation of saturation curvature. Although, as already emphasized, ordinary substances such as water do not exhibit spontaneous saturation, a slight saturation curvature is detectable in very strong applied fields, of the order 100,000 volts/cm. That is to say, the dielectric constant is not quite a linear function of field strength, but instead is given by a formula of the form $\epsilon = \epsilon_0 - aE^2$, the terms beyond E^2 in the series development in E being negligible. If one makes the computation with free rotation and the local field of Lorentz, a ridiculously large negative value of a results. For example, one thus obtains $aE^2/\epsilon_0 = 3.7$ for water in a field of 100,000 volts/cm, whereas actually aE^2/ϵ_0 is only 1.1×10^{-3} according to the measurements of Malsch. Debye finds, however, that with hindered rotation, the calculation yields exactly the right value of a/ϵ_0 to within a percent or so, if the internal field is determined so as to give the right reduction for the dielectric constant in weak fields, i. e. the right value for the main term ϵ_0 .

Despite all these arguments in favor of the use of the hypothesis of hindered rotation to explain away the $4\pi/3$ catastrophe, it appears to the writer that there is one very fundamental objection to this procedure. Namely, the study of discontinuities in specific heats and dielectric constants has usually been regarded as yielding the critical temperatures above which molecules rotate freely, since these discontinuities have been interpreted by Pauling¹⁸ and others as due to the sudden disappearance of hindered rotation when the temperature is raised above a certain critical temperature T_A . If Pauling's viewpoint is accepted, $4\pi/3$ catastrophe is avoided only if his discontinuity temperature T_A is higher than the Curie temperature T_c given by (5). Otherwise as the temperature is reduced from a very high value, a $4\pi/3$ catastrophe will appear before the temperature is lowered sufficiently for hindered rotation to become effective. Actually, however, the condition $T_A > T_c$ is not fulfilled, for example, in the halogen halides, which are, perhaps, the simplest examples of polar molecules. According to Pauling, HCl has $T_A = 100^\circ$ K, and HBr $T_A = 90^\circ$, as these are the critical temperatures of anomalous specific heats and discontinuities in the dielectric constant. On the other hand, one calculates according to (5) that HCl has $T_c = 260^\circ$ and HBr

¹⁸ Pauling, L. Phys. Rev. 38: 430. 1930.

$T_c = 120^\circ$. Thus HCl and HBr should be ferro-electric in the intervals 100–260° and 90–120° respectively, contrary to experiment. Similar difficulties are found in many other molecules. Much of the detailed tabulation by C. P. Smyth¹⁴ of the critical zones of hindered and free rotation for a large variety of substances would have to be rejected if the Debye-Fowler mechanism is invoked to avoid saturation difficulties. Thus it seems necessary to either reject the Debye-Fowler fashion of using free rotation, or else to abandon the Pauling theory of the discontinuities in specific heats, etc., and to attribute the latter to polymorphic transitions. Of course one might assume two critical temperatures for the onset of hindered rotation, but this seems rather an absurdity, as it is hard to see how one can congeal a molecule twice.

Since there is a great deal of evidence in favor of the Pauling-Smyth interpretation of the critical temperatures of free rotation, it becomes of interest to examine whether or not there is some other way than the Debye-Fowler mechanism to avoid the difficulty of the $4\pi/3$ catastrophe. The one outstanding possibility of this kind is that the Lorentz expression (1) for the local field may be grossly incorrect. One has usually regarded the result (1) formula as sacrosanct because it was obtained by the great Lorentz. Naturally his results were correct for the case which he was considering. However, he considered only induced polarization, and so could make his calculation with the harmonic oscillator model. The error has been made by other physicists in seeking to apply his expression for the local field to a case which he did not have in mind, namely the case of polar molecules, where the polarization is due to the orientation of permanent dipoles, rather than to induced dipole-moments. So one may well question whether formula (1) can be applied to the polar media with which we are concerned.

THE ONSAGER FIELD

A completely different expression for the local field has recently been proposed by Onsager.¹⁵ The difference between the Lorentz and Onsager models will be apparent from examination of FIGURE 1. In both of them the given molecule is regarded as at the center of a spherical cavity. In the calculation of Lorentz, the lines of force

¹⁴ Smyth, C. P. Chem. Rev. 19: 329. 1936.

¹⁵ Onsager, L. Jour. Am. Chem. Soc. 58: 1486. 1936. During the printing of the present paper an extension of Onsager's theory which applies particularly well to water has been developed by Kirkwood, J. G. Jour. Chem. Phys. 7: 911. 1939.

are regarded as undeflected by the cavity. That is to say, one congeals the lines of force, and then hollows out the cavity, the argument being that, after all, the molecules are really present instead of being missing. On the other hand, in the Onsager model, the lines of force have cognizance of the existence of the cavity, and are bent by it. However, the "back-action" of the given molecule, which tends to polarize the surrounding medium, etc., cannot be overlooked, and gives rise to what is called the "reaction field" by Onsager, and is indicated by the dotted lines in FIGURE 1. The total field is, of course, the combination of the original field in the absence of the molecule, and of the reaction field, in other words, the sum of the solid and dotted lines in the part of FIGURE 1 illustrating Onsager's model. It is this sum which should be compared with the solid lines in the Lorentz model, which does not attempt to resolve the two parts. Lorentz's calculation covers correctly the special case that the molecular moment is parallel to the applied field E , for his induced dipoles are in this connection typical of any moment aligned along E . Hence when the elementary dipole is parallel to the applied field, the two types of lines of force in the Onsager picture must compound to give straight lines and agreement with Lorentz. Unfortunately, this condition is satisfied only if the volume of the Onsager cavity is taken equal to the molecular volume $1/N$, where N is the number of molecules/cc. It must be regarded as an element of weakness in the Onsager theory that agreement with the correct Lorentz result for the contribution of the induced polarization, or for the special case of parallelism for the permanent moment, is secured only for this particular size for the cavity, which we assume henceforth. In the calculation of Lorentz, on the contrary, the radius of the cavity is arbitrary, and is usually considered so large as to embrace a large number of molecules. In other words, in the Onsager model, the cavity has to be identified physically with the size of the molecules, whereas with Lorentz it was a mathematical fiction which could be arbitrarily chosen.

Clearly in the Onsager mechanism, the reaction field is always parallel to the elementary dipole, and can have no effect on orienting it. Thus for the effective field for determining the spatial distribution of dipoles in the partition function, we should take only the "direct field," represented by the solid lines in the picture of his model. On the other hand, the total field is employed in the usual calculation with the Lorentz model. It is here, according to Onsager, that an

error has been committed in the application of the Lorentz local field to polar molecules, for this field represents essentially the sum of the direct field, and the mean value of the reaction field. So the Lorentz model yields too large an effective orienting field, because it thus includes a part of the reaction field which would correspond to the molecule lifting itself up, or rather orienting itself, by its own bootstraps.

If one neglects the induced polarization, the Onsager model gives the following formula for the dielectric constant:

$$\epsilon = \frac{1}{4} + \frac{3}{4}\psi + \frac{3}{4}\left(1 + \frac{2}{3}\psi + \psi^2\right)^{\frac{1}{2}} \text{ with } \psi = \frac{4\pi N\mu^2}{3kT} \quad (9)$$

The appearance of the radical sign in (9) is, at first sight, rather surprising, since most dielectric formulas do not involve irrationalities. Eq. (9) can be derived by noting that according to Onsager, when there is no induced moment the polarization is

$$P = (N\mu^2/3kT)E_{cen}. \quad (10)$$

Here E_{cen} is the field at the center of a spherical cavity in a medium of dielectric constant ϵ subject to an applied field E , i. e. subject to the boundary condition that the field at a very large distance from the cavity is E . This problem of finding E_{cen} , i. e. of determining the solid

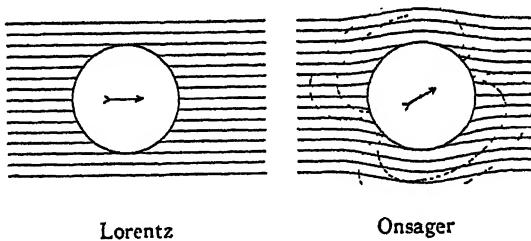


FIGURE 1.

lines of force in the right half of FIGURE 1, is solved in elementary books on electrostatics, and gives

$$E_{cen} = \left(\frac{3\epsilon}{2\epsilon + 1}\right)E \quad (11)$$

Usually in the literature the case is considered that the medium and sphere have dielectric constants unity and ϵ respectively, rather than the converse as we require it, but one readily passes from one case to the other by the substitution $\epsilon \rightarrow 1/\epsilon$. One obtains the result (9) on combining (10) and (11) and remembering that $\epsilon - 1 = 4\pi P/E$.

When one includes also the induced polarization, the formula yielded by the Onsager model is more complicated. Onsager shows¹⁶ that it can be written in the form

$$\frac{\epsilon - 1}{\epsilon + 2} - \frac{4\pi}{3} \left(N\alpha + \frac{N\mu^2}{3kT} \right) = (f - 1) \frac{4\pi N\mu^2}{9kT} \quad (12)$$

with

$$f = \frac{3\epsilon(n^2 + 2)}{(2\epsilon + n^2)(\epsilon + 2)}.$$

Here n is what Onsager calls the "internal index of refraction" and is connected with the cavity radius a and the constant α of induced polarizability, according to the relation $(n^2 - 1)/(n^2 + 2) = a/c^3$. As previously explained, Onsager's formulas reduce properly to Lorentz's in the limiting case of parallelism or of purely induced polarization only if the volume of the cavity equals the atomic volume, so that $4\pi a^3/3 = 1/N$. Then n is identical with the ordinary optical refractive index for infinite wave length, and one has the usual relation

$$(n^2 - 1)/(n^2 + 2) = 4\pi N\alpha/3. \quad (13)$$

With the Onsager model, there is no $4\pi/3$ catastrophe, even with free rotation, for the expression (9) or (12) for the dielectric constant never becomes infinite except at the absolute zero. Instead, isotropic media should never be ferro-electric, in accord with experiment.

Not only does the Onsager model avert a $4\pi/3$ catastrophe, but also it describes the experimental results semi-quantitatively. Onsager shows that if the internal or effective index of refraction n in (12) be given a suitable value, not necessarily, however, in accord with (13), then his expression is substantially the same as Wyman's empirical formula (6). Quite recently Böttcher has analyzed a large amount of experimental data in the light of the Onsager formulas, using the correct value (13) of n . TABLE 2 shows the dipole moments thus calculated by Böttcher for a number of molecules from measurements on pure liquids, while TABLE 3 shows the dipole moments which Böttcher¹⁷ obtains by applying the Onsager relations to ethyl bromide at various temperatures. If the Onsager model were quantitatively accurate, then the dipole moments obtained from the pure liquid in TABLE 2 should be the same as the values obtained from measurements

¹⁶ According to a recent publication of Zakrzewski, K., & Piekara, A. Nature 144: 250. 1939. Eq. (12) is not a correct consequence of Onsager's hypothesis, but we are unable to agree with their contention.

¹⁷ Böttcher, C. J. F. Physica 6: 59. 1939.

TABLE 2
DIPOLE MOMENTS μ IN DEBYE UNITS 10^{-18} E.S.U., CALCULATED FROM ONSGAER'S FORMULA

	calc.	vap.	sol.	calc.	vap.	sol.
nitrobenzene	4.2	4.2	3.9-4.0	CH ₂ Br ₂	1.4	—
nitromethane	3.7	3.4-3.8	3.0-3.1	CH ₃ I ₃	1.1	—
O-nitrotoluene	3.9	3.6	3.7	CHBr ₃	1.0	—
acetone	3.0-3.1	3.0	2.7-2.8	CHCl ₃	1.2-1.3	—
methyl ethyl ketone	3.2	—	2.7-2.8	CH ₂ Cl ₂	1.7	—
acetophenone	3.2	—	2.9-3.0	CH ₃ Cl	1.8	1.7-1.8
aniline	1.5	1.5	1.5-1.6	CH ₂ Br	1.6	1.8
quinoline	2.1	—	2.1-2.2	CH ₃ I	1.3	1.3-1.6
pyridine	2.3	—	2.1-2.2	C ₂ H ₅ Br	1.8-1.9	1.8-2.1
acetonitrile	3.6	3.9	3.1-3.5	C ₂ H ₅ I	1.5-1.7	1.6-2.0
benzonitrile	3.6-3.7	4.4	3.9	C ₆ H ₅ Br	1.4	1.5-1.7
acetaldehyde	2.7	2.7	—	C ₆ H ₅ Cl	1.5-1.6	1.6-1.7
ethyl acetate	1.8	1.8	—	o-C ₆ H ₄ Cl ₂	2.2	—
diethyl ether	1.4-1.5	—	1.1-1.2	m-C ₆ H ₄ Cl ₂	1.4	—
anisole	1.5	1.2	1.2-1.3	ethylene chloride	2.0-2.1	1.2-1.6
acetic acid	1.3-1.7	1.7	—			
water	3.0-3.1	1.8-1.9	1.7-2.0			
ethyl alcohol	2.8-3.1	1.7	1.7			

TABLE 3

DIPOLE MOMENTS FOR ETHYL BROMIDE AT VARIOUS TEMPERATURES CALCULATED FROM ONSAGER'S FORMULA

$\mu \times 10^{18}$	T calc	-90°	-50	-30	0	+30	38.4 (b.p.)
		1.80	1.83	1.85	1.89	1.91	1.95

on vapors or dilute solutions which are included for purposes of comparison and which are presumably correct because our mooted question of dipole-dipole interaction does not enter at low densities. Also in TABLE 3 the moments should be independent of temperature. It is seen that these various tests are not rigorously fulfilled, but all in all, in view of the well-known difficulties attendant to developing a dielectric theory for high concentrations, the situation must be regarded as reasonably satisfactory. It must be mentioned, however, that Böttcher's method of testing the Onsager theory is the most optimistic fashion of procedure. Due to the different ways the constants appear, the percentages of error are greater when the polarization is computed with μ regarded as known, rather than with the reverse scheme used by Böttcher. Even with the latter, the results are rather sensitive to the value employed for n or a . Thus in the case of ethyl bromide (C_2H_5Br) Smyth computes⁹ from Onsager's formula values of the molar polarization P^{OR} defined in (8) which are from 25 to 50 percent larger than the experimental values, and concludes that the agreement is "unsatisfactory" in contrast to the apparently reasonably good accord between the calculated and observed dipole moments for this substance reported by Böttcher (cf. TABLE 2).

When the Onsager field is employed, the difficulty of excessive saturation curvature disappears. If we write $\epsilon = \epsilon_0 - a/E^2$ there is, to be sure, no longer the quantitative agreement between the calculated and observed values of the ratio a/ϵ_0 which Debye computed with his model of hindered rotation. Instead, the calculated value of a/ϵ_0 is too small by a factor of the order 10^{-2} , in marked contrast to the enormous value which the Lorentz hypothesis (1) would yield with free rotation. The absence of quantitative agreement with the Onsager model is not disconcerting, and the exact accord on saturation curvature furnished by Debye's calculation with hindered rotation was probably fortuitous, for any explicit model is an idealized oversimplification. It is gratifying that the computed values of a/ϵ_0 should prove too small with the Onsager hypothesis, for the latter

probably over-accentuates the departures from the Lorentz mechanism (1), and, as we shall see later, the truth probably lies somewhere between the results of Onsager and Lorentz. Thus it appears that study of saturation effects can no longer be regarded as furnishing evidence in favor of the hypothesis of hindered rotation. One can also show¹⁸ that the Onsager model with free rotation yields fully as satisfactory results on the Kerr effect and on dispersion at radio wave-lengths as does the Debye-Fowler mechanism based on hindered rotation, though these phenomena have sometimes been quoted as evidence in favor of the latter hypothesis.¹⁹

EXACT CALCULATION OF THE POLARIZATION AS A SERIES

In view of the above, the Onsager model appears quite as capable of describing the experimental data on dielectric constants as does the Debye-Fowler hindered rotation. If the empirical results consequently do not decide between the two mechanisms, one next asks whether an exact theoretical calculation cannot be made which will decide between the two. Both the Lorentz and Onsager schemes are somewhat phenomenological, and one wonders whether the interactions between the elementary dipoles cannot be treated rigorously rather than by means of a model. In principle this can be done. The problem consists in calculating the partition function

$$Z = \sum_{\lambda} e^{-W_{\lambda}/kT},$$

where the W_{λ} are the characteristic values of the Hamiltonian function

$$H = -E \sum_i \mu_i + \sum_{i>j} \frac{1}{r_{ij}^3} \left[\mu_i \cdot \mu_j - \frac{3(\mu_i \cdot \mathbf{r}_{ij})(\mu_j \cdot \mathbf{r}_{ij})}{r_{ij}^2} \right]. \quad (14)$$

Here the μ is the vector moment of atom i , and the sum is over all the atoms in unit volume. The applied field E is supposed directed along the z axis. The first and second terms of (14) represent respectively the energy due to the applied field, and the coupling between the dipoles. Once the partition function has been calculated, the polarization P can be found by a simple differentiation, *viz.*,

$$P = kT \partial \log Z / \partial E.$$

Unfortunately the determination of the characteristic values W_{λ} constitutes a problem of insoluble complexity, since they are of the order

¹⁸ Cole, R. Jour. Chem. Phys. 6: 385. 1938.

¹⁹ Debye, P., & Bamm, W. Ann. Physik. 23: 28. 1937.

10^{21} in number, owing to the fact that our dynamical system is composed of N molecules. The best that can be done is to expand the partition function in a series of descending powers of kT . Then (14) becomes

$$Z = \eta \left[1 - \frac{\bar{H}}{kT} + \frac{\bar{H^2}}{2k^2 T^2} - \frac{\bar{H^3}}{6k^3 T^3} + \frac{\bar{H^4}}{24k^4 T^4} - \dots \right], \quad (15)$$

where η , the number of states, is a constant independent of E or T , and the bar denotes the quantum-mechanical average. In writing (15) we have made use of the invariance of the diagonal sum, which enables one to express the sum over the various states in terms of an average value which is invariant of the system of quantization and which can be calculated without the necessity of finding the individual characteristic values W . It is for the latter reason that the series method can be employed. However, the labor of calculation increases greatly with the number of terms included. We have succeeded²⁰ in making the computation to H^4 inclusive. This corresponds to only the second power of the dipole-dipole interaction, as Z must be known to E^2 to obtain the part of the polarization linear in the field strength, and so two of the four powers in H^4 must be apportioned to the first rather than the second (dipole-dipole) term in (14). The resulting formula for the polarization proves to be²⁰

$$P = (E/4\pi) \left[\psi + \frac{1}{3} \psi^2 - (Qx/32\pi^2) \psi^3 - \dots \right], \quad (16)$$

where

$$\psi = 4\pi N \mu^2 / 3kT, \quad Q = 2N^{-2} \sum_i r_i^{-6}, \quad x = 1 + [3/8(J^2 + J)], \quad (17)$$

J being the inner quantum number. The departures of x from unity represent a purely quantum-mechanical effect, without particular bearing on the relative correctness of the Lorentz and Onsager models, so we shall henceforth use the value $x = 1$ appropriate to classical theory, which corresponds to the limit $J = \infty$. For purposes of comparison, we may develop the formulas (3) and (9) furnished by the Lorentz and Onsager theories as power series in $\psi = 3T_e/T$. One thus obtains

$$(\text{Lorentz}) \quad P = (E/4\pi) \left(\psi + \frac{1}{3} \psi^2 + \frac{1}{9} \psi^3 + \dots \right), \quad (18)$$

$$(\text{Onsager}) \quad P = (E/4\pi) \left(\psi + \frac{1}{3} \psi^2 - \frac{1}{9} \psi^3 + \dots \right). \quad (19)$$

²⁰ Van Vleck, J. H. Jour. Chem. Phys. 5: 320. 1937.

Since both Onsager and Lorentz envisage a dielectric as a continuum, one is tempted to replace the sum by an integral whose lower limit is the radius of the Onsager cavity having a volume equal to the mean atomic volume. One thus finds

$$Q = 8\pi N^{-1} \int_a^{\infty} r^{-4} dr = 8\pi/3a^3 N = 2(4\pi/3)^2 \quad (20)$$

With the value (20) of Q , the third coefficient in (16) acquires a value exactly equal to that $-1/9$ found in the Onsager formula (19). One is consequently tempted to say that the Onsager model checks with the exact formula (16) through one more term than in the case of the Lorentz model. However, the use of the value (20) of Q appropriate to a continuum is not really allowable. When, instead, one evaluates Q for a face-centered lattice, which is fairly typical of a discrete atomic arrangement, and which is the actual lattice for the solid halogen halides, one finds $Q = 14.4$. The third coefficient in (16) then becomes $-1/22$, about half way between the Lorentz and Onsager values $+1/9$, $-1/9$, though somewhat closer to the latter than to the former. So all that the exact calculation shows is that the truth is somewhere between the Lorentz and Onsager models, as one would have suspected anyway. In fact, existing theory does not appear adequate even to say whether or not dipole-dipole coupling should ever make an isotropic medium have a Curie point and exhibit a state of saturation analogous to ferromagnetism. A complete rather than series calculation would be required to answer this point.

BEARING OF EXPERIMENTS ON MAGNETIC BODIES

Perhaps the reader may say that although the experimental results on dielectric constants, and also the theoretical calculations, are inconclusive as to the existence of a Curie point, the question is settled by the occurrence of isotropic ferromagnetic substances in the related field of magnetism. However, here the ferromagnetic saturation is due to exchange rather than dipole-dipole coupling. The exchange potential has a different structure $-2J\mu_i \cdot \mu_j$ than the dipolar one, given by the second member of (14), and can be shown theoretically to be much more likely than the latter to give a state of spontaneous magnetization. The exchange integrals vary exponentially with distance, rather than as the inverse cube, and so the influence of exchange coupling can be eliminated by using materials of high magnetic dilution. If under these circumstances there is a Curie point, it should

be due to dipolar rather than exchange coupling. However, the resulting magnetic Curie points, T_c , calculated from Eq. (5), if they then exist at all, should be only of the order $.01^\circ$ K. This is to be compared with 10^3 degrees in the electrical case—the difference is due to the great diversity in the magnitude of electrical and magnetic units.

Until recently, Curie temperatures of the order $.01^\circ$ K such as are to be expected in paramagnetic bodies from dipolar in distinction from exchange coupling would have been regarded as far too low to be capable of experimental investigation. With the new method of producing very low temperatures by means of adiabatic demagnetization, however, this is no longer the case. Simon, Kurti and collaborators²¹ indeed find that in iron ammonium alum, a material probably sufficiently dilute that exchange effects may be neglected, a sort of ferromagnetism appears in the neighborhood of $.034^\circ$ K. There they find remanence and hysteresis, though the latter is very feeble, *i. e.* with a small area under the loop. At first sight this discovery seems to indicate that dipole-dipole coupling does lead to a state of ferromagnetism, and that Lorentz is more nearly right than Onsager. Nevertheless, this is not really the case. According to the Lorentz theory, ferromagnetism due to dipolar forces should exist only for specimens which are elongated in the direction of the lines of force. For test bodies which are not extremely prolate, there are demagnetizing corrections which depend on the shape of the test body and express the fact that in the presence of the specimen the field H is not the same as that H_0 which exists in its absence and which is measured experimentally. (The electric analog of demagnetizing conditions does not arise in practice, as one measures directly the drop of potential across condenser plates, and so determines the field actually present inside the body; consequently a $4\pi/3$ catastrophe, if existent, should occur in dielectrics of all shapes.) According to a simple calculation, for a spherical body the demagnetizing and Lorentz dipolar corrections just cancel, so that the effective local field is not $H_0 + 4\pi M/3$ but rather H_0 . Hence in the magnetic case, a $4\pi/3$ catastrophe and a state of spontaneous magnetism should never be found with a spherical body. However, Simon finds his remanence and hysteresis even when a sphere is employed.^{21a} The hysteresis is, to be sure, even feebler

²¹ Kurti, N., Lainé, P., Rollin, B. V., & Simon, F. Compt. Rend. 202:1576. 1936; 204: 675, 754. 1937.

^{21a} In further unpublished work, Ashmead, on the other hand, finds that the saturation anomalies disappear if the ratio of major and minor axes is less than $1.8 : 1.0$. Sauer finds that the critical ratio given by the Lorentz field is $6 : 1$. Sauer, J. A. Phys. Rev. 57: 142. 1940.

than for an elongated specimen, but occurs at approximately the same temperature. Because of this anomalous fact, the ferromagnetic condition found by Simon and collaborators must be of a somewhat pathological variety, perhaps in some way connected with the absence of complete equilibrium at very low temperature. So it cannot be regarded as a confirmation of the use of the Lorentz local field, or as throwing any light on our mooted question of whether dipolar forces lead to a ferromagnetic or ferro-electric state. Temperley and Sauer^{21b} suggest that some of the anomalies at low temperatures may be due to antiferromagnetism rather than ferromagnetism, *i. e.*, to an ordered state in which the dipoles are aligned antiparallel rather than parallel.

INTERPOLATION BETWEEN THE FORMULAS OF LORENTZ AND ONSAGER

Since all the foregoing considerations seems to indicate that the truth lies somewhere between the results of Lorentz and Onsager, it is natural to try to use an interpolation formula which is somewhere in between the two. The Clausius-Mossotti expression (3), based on the Lorentz field, can be obtained from Eq. (12) furnished by the Onsager model by replacing the right hand member of (12) by zero. Hence one possible interpolation scheme consists in multiplying the right side of (12) by a correction factor q which is intermediate between zero and unity, so that (12) becomes

$$\frac{M \epsilon - 1}{\rho \epsilon + 2} - \frac{4\pi L}{3} \left(\alpha + \frac{\mu^2}{3kT} \right) = q(f-1) \frac{4\pi L \mu^2}{9kT}. \quad (21)$$

R. Cole¹⁸ has shown that with certain well-warranted approximations, (21) reduces substantially to the empirical expression (7) of van Arkel and Snoek which represents experimental data so well. The same conclusion, that interpolation between Lorentz and Onsager yields the formula of van Arkel and Snoek, was also obtained simultaneously and independently by Böttcher.²² His way of deriving this result is particularly simple, and is as follows. We have seen that the gist of Onsager's observations is that the Lorentz expression for the local field is in error because it includes the mean value of the reaction field, *i. e.* the average of the projection of the latter in the direction of the applied field E . This average is clearly proportional to the dipole moment μ and to the mean value of the cosine of the

^{21b} Temperley, H. N. V. Proc. Cambridge Phil. Soc. 36: 79. 1940. Sauer, J. A., & Temperley, H. N. V. Proc. Roy. Soc. 176: 203. 1940.

²² Böttcher, G. J. F. Physica 5: 635. 1938.

angle φ between the molecular moment μ and the field E . It can also be shown²⁸ to be proportional to the number N of molecules/cc. Consequently, according to Böttcher, the local field which should be used is

$$E_{local} = E + (4\pi/3)P - R'\mu N \overline{\cos \varphi}, \quad (22)$$

where R' is a constant, which need not necessarily (*i. e.* with the interpolation hypothesis) have as high a value as that corresponding to the Onsager model. According to a well-known formula of Langevin, one has

$$\overline{\cos \varphi} = (\mu/3kT)E_{local}, \quad (23)$$

provided the effective field acting upon the dipole is independent of the latter's orientation, as Böttcher tacitly assumes. Combining (22) and (23), we see that

$$\overline{\cos \varphi} = \frac{\mu}{3kT + R'N\mu^2} \left(E + \frac{4\pi}{3}P \right).$$

The full Lorentz field is considered to act upon the induced polarization. Thus the basic equations differ from those of the conventional theory based upon the Lorentz hypothesis only in that kT is replaced by $kT + \frac{1}{3}R'N\mu^2$ and so the Clausius-Mossotti formula (3) now is modified to become

$$\frac{M\epsilon - 1}{\epsilon + 2} = \frac{4\pi L}{3kT + R'N\mu^2} \left[\alpha + \frac{\mu^2}{3kT + R'N\mu^2} \right]. \quad (24)$$

Eq. (24) is obviously of the same structure as the formula of van Arkel and Snoek given in Eq. (7). The values of R' or of C corresponding to Onsager's hypothesis can be shown to be

$$R' = \frac{4\pi C}{3} = \frac{4\pi}{3} \frac{2\epsilon - 2}{2\epsilon + n^2} \frac{n^2 + 2}{3}. \quad (25)$$

Usually the empirical value of C is somewhat lower than that given by (25). This will be the case if the truth is somewhere between Lorentz and Onsager. Sometimes, however, even the Onsager model gives insufficient reduction in the dielectric constant, and it is necessary to use a value of C larger than (25). An example is furnished by ethyl bromide, where (25) yields $C = 1.32$ and 1.06 at -90° and 30° respectively, whereas Smyth's⁹ experimental values are 1.49 , 1.32 . The corresponding values of the factor q in (21) are 1.05 , 1.09 .

²⁸ For further discussion of this point, which leads back essentially to the Onsager model, see Böttcher.²²

THE LAW OF CORRESPONDING STATES

We have seen that the problem of dipole-dipole interaction, even with free rotation, is one which cannot be solved rigorously. However, even without detailed calculation, there is one very simple result, which follows essentially from dimensional considerations. Namely, if we use "reduced units," *i.e.* express the polarization in terms of its ratio relative to the saturation value $N\mu$, and the temperature in terms of a scale in which the unit is the critical temperature $T_c = 4\pi N\mu^2/9$ (cf. Eq. 5) given by the Lorentz hypothesis (or of any other scale proportional to μ^2), then for any given type of lattice, such as face-centered, body-centered, etc., the curve representing the

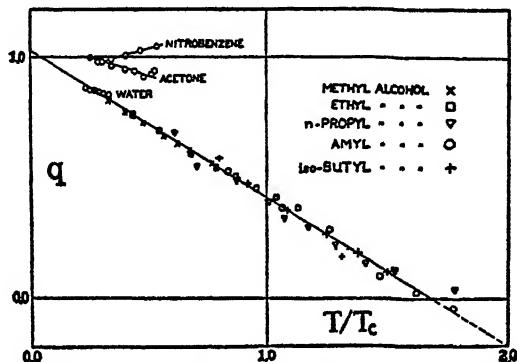
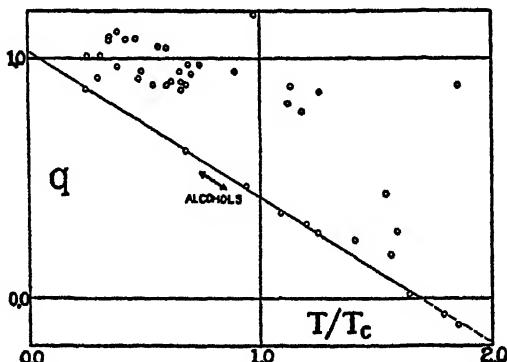


FIGURE 2. The parameter q of Eq. (21) plotted against the reduced temperature T/T_c for the alcohols and water at different temperatures.

polarization as a function of temperature should be the same for all materials. In other words, the functional relation between $P/N\mu$ and T/T_c should be a universal one. There should thus be a law of corresponding states analogous, for instance, to the well-known examples encountered in connection with van der Waals equation, and especially in the Weiss theory of ferromagnetism. (In the latter, the Curie point used in the temperature scale is the real empirical one, whereas ours is a purely hallucinatory one characteristic of the non-existent $4\pi/3$ catastrophe, but this fact has no bearing on the universal functional form.) Another way of expressing the law of corresponding states, used by Cole, is that the interpolation parameter q appearing in (21) should be a universal function of T/T_c . In this connection q is to be regarded as a purely empirical quantity which is so determined as to make (21) satisfied, and which is thus simply a device for recording dielectric measurements. FIGURE 2, taken from

a paper by Cole,¹⁸ shows that the law of corresponding states holds quite well for a series of alcohols, and that q does in general have values intermediate between those zero and unity characteristic of the Onsager and Lorentz theories. Since empirically q is approximately equal to unity at low temperatures, and is very small at high temperatures, it would appear that the Onsager and Lorentz schemes were idealized limits corresponding respectively to very low and very high reduced temperatures, but as yet no theoretical basis has been proposed why this should be so. Furthermore, there are many exceptions to the law of corresponding states. FIGURE 3 shows the results of a large number of substances at room temperatures, for which data are



dipolar orientation. The hypothesis of hindered rotation presents a particular, rather extreme form of such short range forces. The fact that the law of corresponding states holds as well as it does perhaps suggests that most of the departures from the Clausius-Mossotti formula are due simply to the fact that the latter is not a rigorous expression for the results of ordinary dipole-dipole coupling.

SUMMARY

Recently two distinct viewpoints, namely the Debye-Fowler hypothesis of hindered rotation, and the Onsager field, have been proposed which succeed in interpreting a number of phenomena connected with the dielectric constants of liquids and solids. All told, the Onsager theory appears fully as fruitful as does hindered rotation in explaining the various empirical facts, and furthermore, does not run into contradiction with the Pauling explanation of discontinuities in specific heats and dielectric constants in terms of a critical temperature for the disappearance of free rotation. It is doubtful whether dipolar forces can ever lead to a ferro-electric condition in an isotropic medium. In our opinion, the Debye-Fowler hindered rotation has little relation to that of Pauling, and perhaps has reality only to the extent to which it can be regarded as tautological with an accurate representation of dipole-dipole interaction. It is probably because of this tautology that the Debye-Fowler theory has in some ways been quite successful. Thus the distinction between the Onsager field with freely rotating molecules and the Lorentz field with superposed hindered rotation may not be as great as would first appear.

THE LOCAL FIELD IN DIELECTRICS

By JOHN G. KIRKWOOD

From the Baker Laboratory, Cornell University, Ithaca, New York

The interpretation of dielectric polarization from the standpoint of molecular theory demands a knowledge of the average local field acting in the interior of a molecule of a specimen of dielectric, subjected to the action of an external electric field. An approximation to the local field, which has been widely used, is the Lorentz field, $E + 4\pi P/3$. Although the Lorentz field is a fair approximation in non-polar substances, it is entirely inadequate in the case of polar substances, the molecules of which possess permanent electric dipole moments. Even in non-polar substances the Lorentz field is in need of correction, and has proved to be moderately successful only because the departure of the local field from the macroscopic field is relatively small.

We shall first discuss the calculation of the local field in non-polar substances. The problem has been considered from slightly different points of view by Van Vleck¹ and by the writer.² If a slab-shaped specimen of dielectric, bounded by two infinite parallel planes, is subjected to an electric field normal to the plane boundaries, it may be shown that the local field has the form,

$$F = E + 4\pi P - NT_{12} \cdot p_1 \quad (1)$$

$$T_{12} = \frac{1}{R_{12}^3} \left[1 - 3 \frac{R_{12} R_{12}}{R_{12}^2} \right]$$

where E is the macroscopic field, P the average polarization, and N the number of molecules in the specimen. T_{12} is the tensor of dipole-dipole interaction between an arbitrary pair of molecules 1 and 2, and p_1 is the induced electric moment of one member of the pair. The average value, $\bar{T}_{12} \cdot p_1$, is to be calculated by the methods of statistical mechanics as a mean value in a canonical ensemble. In classical statistical mechanics, it becomes an average in the rotational and translational configuration space of the system of N molecules constituting the specimen of dielectric. If $\bar{T}_{12} \cdot p_1$ is approximated by $\bar{T}_{12} \cdot \bar{p}_1$, it is found to be equal to $-8\pi P/3N$ in isotropic fluid dielectrics, and the local field of Eq. (1) reduces to the Lorentz field, $E + 4\pi P/3$. The

¹ Van Vleck, J. H. *Jour. Chem. Phys.* 5: 556. 1937.

² Kirkwood, J. G. *Jour. Chem. Phys.* 4: 592. 1936.

deviation from the Lorentz field in non-polar fluid dielectrics is therefore equal to $N[\bar{\mathbf{T}}_{12} \cdot \mathbf{p}_1 - \bar{\mathbf{T}}_{12} \cdot \mathbf{p}_1]$. A part of this term arises from fluctuations in the relative configuration of the molecular centers, in the course of thermal motion. If the molecules are non-spherical and optically anisotropic, fluctuations in orientational configuration also contribute.⁸ Mathematical difficulties make it impracticable to calculate $\bar{\mathbf{T}}_{12} \cdot \mathbf{p}_1 - \bar{\mathbf{T}}_{12} \cdot \mathbf{p}_1$ in closed form. Calculation as power series in the density of the fluid is, nevertheless, possible. In this manner an expansion of the dielectric constant ϵ of the fluid in a power series in the reciprocal of the molal volume v is obtained, which is analogous to the virial expansion of the equation of state of a gas.

$$\begin{aligned}\frac{\epsilon - 1}{3} &= \frac{P_o}{v} \left[1 + \left(1 + \gamma + \sigma \right) \frac{P_o}{v} + \frac{1}{16} \left(\frac{P_o}{v} \right)^2 + \dots \right] \\ P_o &= \frac{4\pi N\alpha}{3} \quad \gamma = \frac{P_o}{b} (1 + a/3bRT) \\ \sigma &= \frac{\alpha_1 - \alpha_2}{\alpha} \left[\left(\frac{1 - e^2}{e^3} \right) \sin h^{-1} \left(\frac{e}{(1 - e^2)^{1/2}} \right) - \frac{1 - e^2}{e^2} - \frac{1}{3} \right]\end{aligned}\quad (2)$$

where α is the mean polarizability of a molecule, and a and b are the constants of the Van der Waals equation of state. The influence of translational fluctuations on the local field is contained in γ . The rotational fluctuation term σ is given only for the special case of a prolate ellipsoidal cavity of eccentricity e , from which an arbitrary molecule excludes the centers of its neighbors by the action of intermolecular repulsive forces. The quantities α_2 and α_1 are the lateral and longitudinal principal polarizabilities of the anisotropic molecules. The corresponding expansion of the Clausius-Mossotti formula, based on the Lorentz field is,

$$\frac{\epsilon - 1}{3} = \frac{P_o}{v} \left[1 + \frac{P_o}{v} + \left(\frac{P_o}{v} \right)^2 + \dots \right] \quad (3)$$

The sum, $\gamma + \sigma$, is in general of the magnitude of 0.1. A comparison of Eqs. (2) and (3) shows that the Lorentz field leads to an error of about ten percent in the coefficient of $1/v^2$, and to an error of about 1600% in the coefficient of $1/v^3$. Fortunately, the latter term is usually rather small even at liquid densities. It should be remarked that the theory has been developed on the assumption that the molecular polarizability α is independent of the density of the fluid. This

⁸ See Raman, C. V., & Krishnan, K. S. Proc. Roy. Soc. A 117: 589. 1927.

is not necessarily true, and dependence of α on the density might introduce additional terms in the coefficients of the power series of Eq. (2).

The deficiencies of the Lorentz approximation to the local field are most spectacularly exhibited in polar liquids. It leads to the prediction of electric Curie points in liquids at high temperatures, although such Curie points have never been observed. They can be avoided only by abandoning the Lorentz field or by assigning to the molecules of polar liquids dipole moments of unreasonable magnitude. Further evidence against the Lorentz field is provided by the work of Wyman on the dependence of the dielectric constant of polar mixtures on composition.⁴ Attempts have been made to salvage the Lorentz field by Debye⁵ and Fowler⁶ by bringing into consideration hindered rotation of a molecule relative to its environment in the liquid. While it is clear that hindered rotation must play a role in the dielectric polarization of polar liquids, it is certainly responsible for large departures from the Lorentz field. The hypothesis should therefore not be used to supplement the Lorentz field, with which it is really incompatible, but should be introduced for the purpose of correcting it. An entirely rigorous formulation of the problem, in which deviations from the Lorentz field due to hindered rotation are implicit, has been given by Van Vleck.⁷ Unfortunately, the method of calculation which he proposes becomes prohibitively difficult in condensed phases. The most successful calculation of the local field in polar liquids is due to Onsager.⁸ He treats the molecule as a real cavity in a statistical continuum of uniform dielectric constant equal to that of the liquid in bulk. On the basis of this model, the electrostatic theory of continuous media leads at once to a simple expression for the local field and average torque effective in orienting a dipole molecule relative to an external field. The Onsager theory fails to be entirely exact because of the assumption of a uniform local dielectric constant identical with that of the fluid in bulk.

It is possible to generalize the Onsager theory in a form which is entirely rigorous within the limits of applicability of classical statistical mechanics.⁹ If optical contributions to the polarization are neglected, the theory leads to the following expression for the dielectric constant of a polar liquid,

⁴ Wyman, J. *Jour. Am. Chem. Soc.* **58**: 1482. 1936.

⁵ Debye, P. *Physikal. Zeit.* **36**: 100. 1935.

⁶ Fowler, R. H. *Proc. Roy. Soc. London.* **149A**: 1. 1935.

⁷ Van Vleck, J. H. *Jour. Chem. Phys.* **5**: 556. 1937.

⁸ Onsager, L. *Jour. Am. Chem. Soc.* **56**: 1486. 1934.

⁹ Kirkwood, J. G. *Jour. Chem. Phys.* **7**: 911. 1939.

$$\frac{\epsilon - 1}{3} = \frac{3\epsilon}{2\epsilon + 1} \frac{P_o}{v}$$

$$P_o = \frac{4\pi N}{9kT} \mu \cdot \bar{\mu} \quad (4)$$

where μ is the molecular dipole moment and $\bar{\mu}$ is the total moment of a molecule and that which it induces in its surroundings within a sphere of radius r_0 , exterior to which the local dielectric constant is effectively equal to the macroscopic dielectric constant of the liquid. A distance r_0 of molecular dimensions, let us say a thousand molecular diameters, certainly exists, beyond which the error in approximating the local by the macroscopic dielectric constant is negligible. If r_0 is assumed to be equal to the molecular diameter, μ reduces to μ , and the Onsager theory is obtained. The next stage of approximation consists in assuming r_0 to be the radius of the smallest sphere containing a molecule and its first shell of neighbors. In this approximation the product $\mu \cdot \bar{\mu}$ becomes,

$$\mu \cdot \bar{\mu} = \mu^2 [1 + z \overline{\cos \gamma}] \quad (5)$$

where z is the average coordination number and $\overline{\cos \gamma}$ is the mean value of the cosine of the angle between the dipole moments of an arbitrary pair of neighboring molecules in the liquid. When the dielectric constant is large relative to unity, Eqs. (4) and (5) lead to the approximate expression,

$$\epsilon - 1 = \frac{2\pi N}{3v} \frac{\mu^2}{kT} (1 + z \overline{\cos \gamma}) \quad (6)$$

The role of hindered rotation in dielectric polarization is clearly exhibited by Eq. (6). Hindered relative rotation of neighboring molecules is responsible for a correlation between their orientations, which may lead to a non-vanishing value of $\overline{\cos \gamma}$. The mean torque hindering the relative rotation of neighbors may be due, in part, to electrostatic dipole-dipole coupling and, in part, to other intermolecular forces. The mean value of $\overline{\cos \gamma}$ may be computed from the potential of mean torque W_0 ,

$$\overline{\cos \gamma} = \int \int \cos \gamma e^{-W_0/kT} d\omega_1 d\omega_2 \quad (7)$$

where the integration extends over all orientations of both molecules of the pair. The calculation of W_0 itself is a problem in statistical mechanics, which has not yet been satisfactorily treated due to purely

mathematical difficulties. Although W_0 is almost certainly a more complicated function of the molecular orientations, a formal calculation of $\cos \gamma$ may be made, if W_0 is assumed to depend only on the angle γ between a pair of axes fixed in the two molecules. We shall further suppose that W_0 can be adequately approximated by a two-term Fourier series,

$$-W_0 = w_0 + w_1 \cos \gamma + \dots \quad (8)$$

where w_0 provides for the normalization of the distribution function, and $2w_1$ is the average work necessary to turn a pair of neighboring molecules from an antiparallel to a parallel alignment of the axes of intermolecular force. Eqs. (7) and (8) lead to the result,

$$\overline{\cos \gamma} = \cos^2 \beta L(w_1/kT) \quad (9)$$

where $L(x)$ is the Langevin function and β is the angle between the dipole moment and the axis of intermolecular force of a molecule. From the experimental value of the dielectric constant, the molecular dipole moment μ and the average coordination number z , it is possible to calculate w_1 in a manner analogous to that of the Debye hindered rotation theory. Since the assumptions concerning the form of the torque potential W_0 can have only a rough correspondence to the real situation, the results of such calculations cannot be taken too literally. Nevertheless, they should provide a qualitative measure of the degree of hindered rotation of neighboring molecules relative to each other in polar liquids.

In some liquids, for example water, hindrance is so strong that a molecule and its first shell of neighbors may be treated as a quasi-rigid structure, and a detailed knowledge of the hindering torque is unnecessary. If we assume a modified Bernal-Fowler¹⁰ structure for the liquid and a bond angle of 100° for the water molecule, the experimental value, 78, of the dielectric constant of liquid water at 25° C is obtained from Eq. (6) with a dipole moment in the liquid 26% greater than that of a water molecule in the vapor. Homogeneous polarization of a molecule by the dipole field of its neighbors is sufficient to account for a 16% increase in moment. With a dipole moment 1.16 times that of a molecule in the vapor, a dielectric constant of 67 is obtained from Eq. (6). This is to be regarded as a rather satisfactory agreement with experiment, since the dielectric constant is very sensitive to the exact structure of a molecule and its first coordination shell, for which we have made some rather rough assumptions of the Bernal-Fowler type.

¹⁰ Bernal, J. D., & Fowler, R. H. Jour. Chem. Phys. 1: 515. 1933.

Although the theory which has been described does not completely solve all problems relating to the dielectric polarization of polar liquids, it reduces them to a problem of somewhat lower order of complexity, namely the calculation of $\bar{\mu}$, the sum of the dipole moment of a molecule and the moment which it induces in its immediate environment in the liquid. To carry out this calculation, we must as yet resort to approximate methods.

THE DIELECTRIC ANOMALIES OF ROCHELLE SALT

BY HANS MUELLER

From the George Eastman Research Laboratories, Massachusetts Institute of Technology, Cambridge, Massachusetts

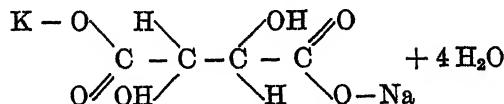
Until some ten years ago the theories of the dielectric and magnetic properties presented a curious dilemma. Langevin's theory of paramagnetism, on one hand, could not account for the existence of ferromagnetic substances while on the other hand the analogous theory for dielectrics, *i. e.* Debye's theory of polar molecules in its original form, predicted that most polar liquids should be "ferroelectric." According to this theory these liquids (or small regions thereof) should become spontaneously polarized below a critical temperature and their dielectric properties should be analogous to the magnetic properties of iron. Such ferroelectric liquids have never been found.

The theoretical developments of recent years have been successful in removing to a large extent these contradictions between theory and observation. Wave mechanics has solved the puzzle of ferromagnetism. As Professor Van Vleck has discussed in the introductory lecture, two modifications of Debye's theory have been proposed. One theory assumes hindered rotation of the dipoles, the other replaces the Clausius-Mosotti hypothesis by Onsager's interaction theory. Both theories are able to explain why liquids do not become ferroelectric.

In contrast to this theoretical work the experimental physicists have succeeded in discovering substances with ferroelectric properties. Since they are found only in certain crystals, this discovery is not contrary to the recent theories. It can be shown that in an anisotropic arrangement of dipoles the conditions are more favorable for a spontaneous polarization to occur. On the other hand we are not justified in considering the discovery of ferroelectricity as a confirmation of any of the present theories. It is indeed doubtful whether the phenomenon is caused by the orientation of dipoles.

THE FERROELECTRIC CRYSTALS

Ferroelectricity has been found in two groups of crystals. The best known representant of the first group is Rochelle salt (Seignette salt). It is the tetrahydrate of Potassium-Sodium-Tartrate.



The isomorphous Ta-Na and Rb-Na salts also are ferroelectric, but the NH₄-Na-Tartrate, though it is also an isomorph, has normal dielectric properties. Rochelle salt has been studied extensively because large uniform crystals can easily be grown. The crystals belong to the ortho-rhombic hemiedric (sphenoidal) class, they are transparent, colorless and have a small biaxial birefringence. The elementary lattice cell of Rochelle salt has the dimensions¹ $a = 11.9 \text{ \AA}$, $b = 14.3 \text{ \AA}$, $c = 6.2 \text{ \AA}$ and contains four molecules. The space group is V_8 , but the lattice structure has not been determined. The ferroelectric properties occur only for electric fields in direction of the a axis, and only within a limited temperature range between the "upper" Curie point at 23.7° C and the "lower" Curie point at -18° C . For fields parallel to the b or c axes the dielectric properties are normal ($\epsilon_u = 12.5$, $\epsilon_s = 10.2$ at 30° C).

To the second ferroelectric group belong the crystals of Dihydrogen-potassium-phosphate H_2KPO_4 and Arsenate H_2KAsO_4 . The isomorphous Ammonium salts $H_2NH_4PO_4$ and $H_2NH_4AsO_4$ probably also are ferroelectric, but the evidence is not complete. These crystals become ferroelectric at temperatures below $-120^\circ C$. The existence of a lower Curie point is doubtful. The spontaneous polarization is in the direction of the tetragonal *c* axis. For fields parallel to the *a* or *b* axis the dielectric behavior is normal, but the dielectric constant changes at the Curie point. The lattice structure of H_2KPO_4 has been determined by West² and is shown in FIGURE 1. The most interesting features of this structure are the hydrogen bonds between neighboring PO_4 groups. The crystal contains no water molecules or other dipoles. If the atomic mechanism responsible for ferroelectricity is the same in both crystal groups (there is no evidence either for or against this assumption) it can therefore not involve rotating dipoles. It seems more likely that the hydrogen bonds play an essential role in this mechanism.

Both groups of crystals have the property of piezoelectricity. An electric field parallel to any of the crystallographic axes of Rochelle salt or to the *c* axis of H_2KPO_4 produces a reversible deformation which in both crystals is a shearing strain in the plane normal to

¹ Krutter, H. M., & Warren, B. E. Phys. Rev. 43: 500. 1933.

² West, J. Zeit. Kristallographie. 74: 306. 1930.

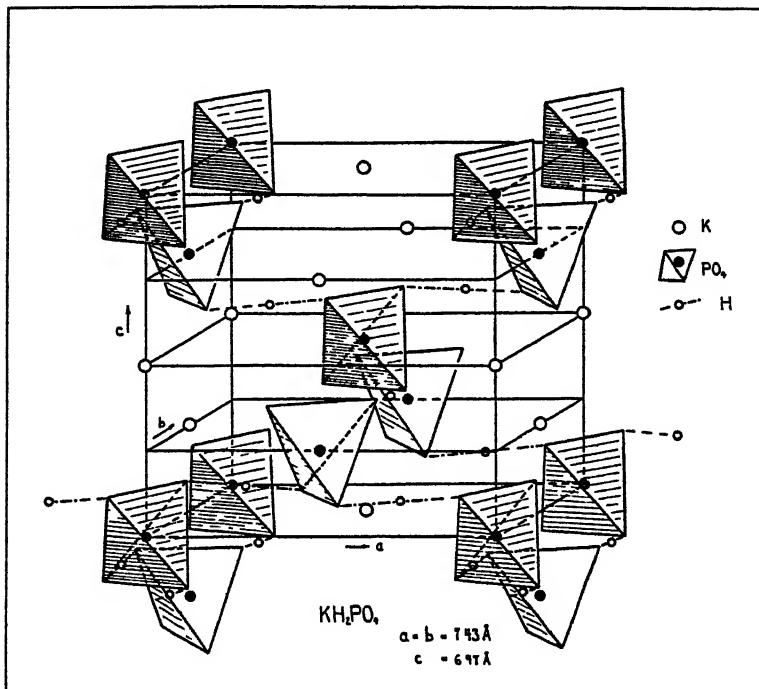


FIGURE 1. Elementary cell of crystal lattice of H_2KPO_4 .

the field. For fields in the ferroelectric direction the piezoelectric constant is very large, as much as a few hundred times larger than in other crystals. The symmetry of both crystalline groups excludes the occurrence of pyroelectric phenomena.

HISTORY AND THEORIES OF FERROELECTRICITY

Cady³ and Anderson⁴ seem to have been the first to observe the anomalous dielectric behavior of Rochelle salt, but the credit for discovering its ferroelectric properties in 1921 goes to Valasek.⁵ The second group of ferroelectric crystals was discovered in 1935 by Busch and Scherrer.⁶ Valasek was the first to observe dielectric hysteresis and saturation and to recognize the analogy with the magnetic properties of ferromagnetic substances. There exists, how-

³ Cady, W. G. Rep. Nat. Res. Council. 1918.

⁴ Anderson, J. A. Rep. Nat. Res. Council. 1918.

⁵ Valasek, J. Phys. Rev. 17: 475. 1921; 19: 478. 1922; 20: 639. 1922; 24: 560. 1924.

⁶ Busch, G., & Scherrer, P. Naturwiss. 23: 737. 1935.

ever, an earlier investigation of Pockels⁷ which came very close to making the discovery in 1893. Pockels was studying the relation between the piezoelectric, electro-optical and photoelastic constants of crystals. In the course of this work he found the large piezoelectric and the linear electro-optical effects in Rochelle salt. In addition Pockels detected a new electro-optical effect which has not been observed in any other crystal.

The linear electro-optical effect is common to all piezoelectric crystals. It manifests itself in a change of the birefringence proportional to the applied electric field. The symmetry properties of Rochelle salt require that this linear effect cannot influence the propagation of a light wave which passes in the direction of a crystallographic axis. Pockels found, however, that the field altered the birefringence even when the light beam was parallel to an axis. Since for these directions the change of double refraction could not be reversed by reversing the field Pockels called the new phenomenon a quadratic electro-optical effect. It is similar to the well known electro-optical Kerr effect in liquids, though it manifests itself in a somewhat different way. In liquids an electric field creates a birefringence proportional to the square of the field strength, in Rochelle salt it produces a change of the already existing natural double refraction. These changes are much larger (1000 to 100000 times) than those produced by the Kerr effect in liquids, but though they are not reversible they usually are not proportional to E^2 .

Pockels realized that the anomalous magnitude of the piezoelectric and of both electro-optical effects might be the result of an abnormal dielectric behavior, but he decided against such a correlation because, some years previous, Borel had measured the dielectric constants of Rochelle salt by a method employing very short Hertzian waves and had reported quite normal values of between 4 to 6 for all crystallographic directions. Instead Pockels suggested that the Kerr effect was due to some new kind of conduction process, and he produced a series of curious facts which, seemingly, supported his point of view. We know now that, while his observations were perfectly correct, the interpretation was wrong. Pockels overlooked that the dielectric properties may be quite different for static fields, which were employed in his researches, from those for high frequency fields. It is most probable that, had Pockels realized this fact, the phenomenon of ferroelectricity might have been discovered 30 years earlier

⁷ Pockels, F. Abh. Ges. Wiss. Gottingen 39: 161. 1893.

and it is interesting to speculate how this would have influenced the history of dielectric research.

The significance of Valasek's discovery was not realized for a number of years. This was partly due to the fact that he encountered great difficulties in securing reproducible data and partly because he suggested no explanation of the phenomenon. Two causes were mainly responsible for a renewed interest in Rochelle salt during the last ten years. One cause is the development of technical application of this substance. Its large piezoelectric effect is utilized for the transformation of sound energy into electrical energy or vice versa in crystal microphones, loudspeaker units, phonograph pickups and recorders and in oscilloscopes. As a result of this development excellent crystals, cut in any desirable shape, have become commercially available to the investigators.⁸ The second impetus was given by the work of Kurtschatow⁹ and his collaborators. This group of Russian physicists proposed the first theory of ferroelectricity. They assume that the phenomenon is due to the orientation of the polar molecules of the water of crystallization in Rochelle salt. In a manner analogous to Weiss' interpretation of ferromagnetism they ascribe ferroelectricity as due to a spontaneous orientation of these dipoles. This occurs at the upper Curie point where the energy of the Lorentz-Lorenz interaction between the dipoles becomes larger than the energy of the temperature motion. The disappearance of ferroelectricity at the lower Curie point has been interpreted by Fowler¹⁰ as due to a gradual freezing in of the rotating dipoles. The assumption of freely rotating water molecules is supported by the fact that Rochelle salt is strongly efflorescent. At room temperature the water molecules are therefore very loosely bound to the framework of the crystal. On the other hand this theory cannot explain why the dielectric properties in Rochelle salt are anomalous only for fields in the *a* direction and it neglects the important interaction between dielectric polarization and piezoelectric deformation. Most investigators¹¹ are now of the opinion that this theory needs further elaboration or a modified foundation.

The origin of ferroelectricity must involve some kind of cooperative or autocatalytic phenomenon. A spontaneous polarization can be created if the polarization of one atom or the orientation of a dipole

⁸ Supplied by the Brush Development Co., Cleveland, Ohio

⁹ Kurtschatow, I. V. Seignette-Electricity, in Russian. Moscow. 1933. French Trans. entitled: Le Champ Moléculair Dans Les Diélectriques. Paris. 1936.

¹⁰ Fowler, R. H. Proc. Roy. Soc. 149A: 1. 1935.

¹¹ Scherrer, P. Zeits. Elektrochem. 45: 171. 1939.

influences the neighboring atom to become polarized in the same direction. The Lorentz-Lorenz theory of the inner field, as used in Weiss' theory, provides such a cooperative action. Cady¹² has suggested another type of autocatalytic process involving the piezoelectric properties of the crystals. Due to secondary effects the dielectric constant of piezoelectric crystal is larger when the crystal is permitted to deform freely than when the deformation is suppressed.¹³ If an electric field is acting on a "free" crystal its deformation creates an additional polarization which in turn further increases its deformation and so on, and it is easily seen how such an autocatalytic interaction eventually can create a spontaneous polarization in some crystals. The experimental evidence shows that this interaction between mechanical deformation and electric polarization plays an important role in the behavior of Rochelle salt.¹⁴ If the piezoelectric deformations are suppressed in a crystal its ferroelectric properties are much less pronounced and the ferroelectric phenomena would most probably disappear if the deformations could be completely avoided. It is difficult to accomplish this because the crystals usually crack during the experiment.

This fact plays an important role in all technical applications of Rochelle salt. The crystals are always used in pairs in such a way that one crystal constrains the other. This procedure reduces somewhat the sensitivity of the various crystal devices without reducing their efficiency (hysteresis losses are greatly reduced) but it has the important advantage of eliminating the strong temperature dependence of the ferroelectric properties of free crystals.

It is probably not accidental that, so far, ferroelectricity has been found only in crystals which also are strongly piezoelectric. Future research will have to decide whether or not this factor is essential, and if so, whether the large piezoelectric constants arise as a result of the orientation of dipoles or of peculiar properties of the hydrogen bonds.

Von Jaffe¹⁵ has discussed a somewhat different aspect of the ferroelectric problem. He has pointed out that at the transition temperatures the orthorhombic structure of Rochelle salt undergoes a polymorphous transition and that in the ferroelectric state the crystals are monoclinic hemimorphous. This conclusion is undoubtedly cor-

¹² Cady, W. G. Phys. Rev. 33: 278. 1929. See also Reference 20.

¹³ Cady, W. G. Am. Physics Teacher. 6: 227. 1938.

¹⁴ Sawyer, C. B., & Tower, C. H. Phys. Rev. 35: 269. 1930.

¹⁵ Von Jaffe, H. Phys. Rev. 51: 43. 1937.

rect. In the ferroelectric state the spontaneous polarization creates a spontaneous piezoelectric deformation which alters the symmetry properties of the crystal. By the same argument we should consider ferromagnetic iron to be tetragonal and not cubic, because the unit cell is deformed through magnetostriction. This point of view involves the danger of confusing cause and effect, it seems to emphasize the deformation as a primary cause and the polarization as secondary effect, while the other theories prefer the opposite. The importance of von Jaffe's contribution lies in the fact that it explains many ferroelectric phenomena (as *e. g.* the infinite dielectric constant and the change of specific heat at the Curie point, the pyroelectric and electrocaloric effects), on the basis of fundamental laws of crystal physics, without resorting to any special hypothesis.

THE PHENOMENOLOGICAL THEORY

Although we have no adequate atomic theory of ferroelectricity it is easy to formulate a simple phenomenological theory. The autocatalytic nature of the phenomenon suggests the assumption of a functional relationship of the form $P = \varphi(E, P)$. For small fields E and small polarizations P this leads to the first approximation

$$P = a(E + fP) \quad (1)$$

where a and f are positive constants which may vary with the temperature T . (1) is the most simple but not necessarily the correct relationship. For H_2KPO_4 Busch¹⁶ has proposed a first approximation of the form $P = a(E + f' P' + f'' P'') = P' + P''$, because it may be argued that the polarization consists of various parts (atomic, lattice, polar) and they will play different roles in the cooperative action. From the simple law (1) it follows that the susceptibility κ_0 for small fields is $\kappa_0 = P/E = a/(1 - af)$, or if we substitute $af = \Theta/T$

$$\frac{1}{\kappa_0} = \frac{f}{\Theta}(T - \Theta) \quad (2)$$

This relation is analogous to the Curie-Weiss law for the paramagnetic susceptibility. Ferroelectricity sets in when the susceptibility becomes infinitely large, *i. e.* when the temperature T reaches the "Curie-temperature Θ ." Since Rochelle salt has two Curie points we must assume that the Curie temperature is not a constant, as in Weiss' theory, but is a function $\Theta(T)$ of the temperature. There must be

¹⁶ Busch, G. Helv. Phys. Acta. 11:269. 1938.

two solutions to the equation $\Theta(T) = T$ which determines the Curie points T_c and $T_{c'}$. For H_2KPO_4 , where the existence of a lower Curie point is very doubtful, Θ might be a constant. The second parameter f is generally assumed to be a constant. No reliable method has as yet been found to verify this assumption for Rochelle salt, and the absolute value of f is not known. If one chooses an arbitrary constant f value the course of the $\Theta(T)$ curve can be calculated with Equation (2) from the observed susceptibilities for small fields. The

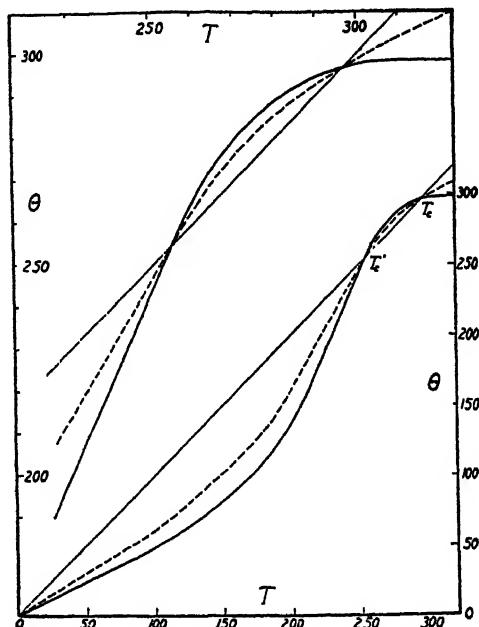


FIGURE 2. The Curie temperature of Rochelle salt, calculated from the susceptibility data for $f = 2.19$ and for $f = 4\pi/3$ (dotted curve).

results of such a calculation, for $f = 2.19$ and $f = 4\pi/3$, are shown in FIGURE 2. From these curves it follows that if f were smaller than 2.19 the $\Theta(T)$ curve would show a maximum. Since this is improbable one infers that f cannot differ greatly from $4\pi/3$. This result agrees, of course, with the prediction of the Lorentz-Lorenz theory of the inner field. Nevertheless it cannot be interpreted as a proof of the dipole hypothesis, because Cady's theory also leads to an f value of this order of magnitude.¹⁷

¹⁷ Cady's theory gives $\kappa_0 = \kappa_1 + \epsilon_1 d_{14}$, where κ_1 is the susceptibility of the constrained crystal. From $P = \kappa_0 E$ it follows $P = \kappa_1 \left(E + \frac{\epsilon_1 d_{14}}{\kappa_0 \kappa_1} P \right)$ and hence by comparison with (1)

On the basis of Fowler's theory $\Theta(T)$ is proportional to the number of effectively free dipoles. This quantity increases continuously with the temperature and may reach asymptotically a constant value near the melting point where all dipoles become free. The experimental results do not support the assumption, made by some writers, that the dipoles are suddenly frozen in at the lower Curie point. This temperature cannot be identified with a "Pauling temperature." The process of "freezing in" occurs gradually, and even far below the ferroelectric temperature range a large number of dipoles must be free to account for the large dielectric constant at low temperatures.

According to the $\Theta(T)$ curves in FIGURE 2 the occurrence of ferroelectricity in Rochelle salt is somewhat accidental. It seems likely that the dielectric susceptibility of other crystals might furnish similar $\Theta(T)$ curves, but only a very few become ferroelectric because these curves usually have no intersect $\Theta = T$. In Rochelle salt the temperature range where $\Theta > T$ is narrow and $(\Theta - T)$ is always small. If its $\Theta(T)$ curve would be only slightly different Rochelle salt would not become ferroelectric. We can therefore expect that slight changes of the structure of this crystal will produce appreciable shifts of the Curie points and may eventually suppress the occurrence of ferroelectricity.

This expectation has been verified in a number of different ways. It has been found^{19, 18} that hydrostatic pressure raises both Curie-points by several degrees. In Rochelle salt with heavy water¹⁹ the two Curie points are farther apart (34.5° and -24° C.). In solid solutions of NH₄-Na-Tartrate in Rochelle salt⁹ the Curie points are closer together, they merge and disappear when more than 3% of the K ions are replaced by NH₄ groups. For molar concentrations of NH₄ up to 20% these mixed crystals are not ferroelectric, but if still more K is replaced by NH₄ they show again a single Curie point below -70° . This point shifts rapidly to lower temperatures with increasing NH₄ concentration and the pure NH₄-Na-Tartrate is again non-ferroelectric.

Since the difference $(\Theta - T)$ is never more than about 10° the dielectric properties of Rochelle salt should not be compared with the magnetic properties of iron at room temperature, but rather with

$f = \frac{\kappa_1 - \kappa_2}{\kappa_2 \kappa_1}$, or since $\omega \gg \kappa_1$, $f = \frac{1}{\kappa_1} - \frac{4\pi}{\kappa_1 - 1}$. The dielectric constant κ_1 for a constrained crystal is small, but its exact value is not known.

¹⁹ Bankcroft, D. Phys. Rev. 53: 587. 1938.

¹⁸ Holden, A. N., Kohman, G. T., Mason, W. P., & Morgan, S. O. Phys. Rev. 56: 378. 1939. Hablitzel, J. Helv. Phys. Acta. 12: 278. 1939.

those of ferromagnetic substances at temperatures near their respective Curie points. For temperatures near the Curie points the polarization becomes extremely large and the first approximation (1) is no longer adequate. Since $(\Theta - T)$ is limited to small values it seems appropriate to try a second approximation of the form

$$P = \frac{\Theta}{fT} (E + fP) - \beta(E + fP)^3 \quad (3)$$

This again is the most simple of all possible generalizations of Equation (1). It introduces only one new parameter β , which is assumed to be independent of temperature. Equation (3) is based on the analogy with Weiss' theory but its justification is largely empirical. (3) implies that the ferroelectric properties depend primarily on the inner field $F = E + fP$. In accordance with this postulate it is natural to assume that all other properties of ferroelectrics depend on F in the same manner as they depend on E in ordinary dielectrics. Thus we postulate for the piezoelectric deformation

$$y_z = d_{14}F \quad (4)$$

and for the birefringence Δ of the Kerr effect

$$\Delta = \varphi F^2 \quad (5)$$

In what follows we shall attempt to show that Equations (3), (4), and (5), together with the empirically determined function $\Theta(T)$, are sufficient to account in a quantitative, or at least semi-quantitative manner, for all electrical, optical and caloric data on Rochelle salt.

EXPERIMENTAL DIFFICULTIES

Before discussing the experimental results we wish to call attention to some of the peculiar experimental difficulties which are encountered in the investigation of Rochelle salt. Most ferroelectric effects are large enough to be observed by very simple methods, but it is difficult to obtain reproducible and consistent data. Uniform crystals can easily be grown or bought and even poor samples can be improved by heating them to a few degrees below the melting point at 54° C. Efflorescence of the crystals can be prevented by controlling the humidity or by coating it with a dilute solution of balsam in xylene. The large temperature variation of all properties and the low heat conductivity of the material call for very careful temperature control. The electrical conductivity, though small, is a disturbing factor particularly in measurements with static fields. The largest errors arise

from the fact that it is practically impossible to keep the crystal free from any constraints. They can be minimized by using electrodes of graphite or thin foil and by suspending the crystal by flexible leads. The electrodes should cover the entire crystal surface because if the edges are not in the electric field the outer parts of the crystal prevent the deformation of the interior. The method of guard rings for dielectric measurements can therefore not be used and is also not necessary, because the dielectric constant ϵ is so large that very few lines of force pass outside the crystal. Due to the large value of ϵ it is necessary that the electrodes be in direct contact with the crystal. Even a very thin layer of air or glue between electrode and crystal reduces the potential drop in the crystal to a fraction of that across the electrodes. This fact also makes it impossible to observe the dielectric anomalies in a powder of Rochelle salt.

In static measurements the results depend on the time during which the field has been acting. These "creep" phenomena are due either to electric conduction or it may be that processes with a large relaxation time are involved.²⁰ Dynamic a. c. methods usually give reproducible results if the frequency is sufficiently high (over 100 cycles/sec). In experiments, where these methods cannot be used, as *e. g.* in the electro-optical investigations, the influence of the time factor can be reduced by repeating every reading a large number of times and by short circuiting the crystal for some minutes after every reading. Successive observations are made with the field reversed and the time of observation is reduced to a minimum.

THE DIELECTRIC PROPERTIES OF ROCHELLE SALT

The dielectric measurements depend on the intensity of the electric field. Measurements in weak fields (less than 10 Volts/cm) are performed with the capacitance bridge or by the resonance method. They furnish κ_0 . Excepting frequencies below 100 cycles/sec and regions where the frequency is near a mechanical resonance frequency of the crystal the variation of κ_0 with frequency is small up to 60 Megacycles. FIGURES 3 and 4 give the dependence of $1/\kappa_0$ on temperature. The $\Theta(T)$ curve in FIGURE 2 for temperatures above the upper and below the lower Curie point have been calculated from these data. Within small temperature ranges the $1/\kappa_0$ curve can be approximated by a Curie-Weiss law $1/\kappa_0 = (t - t_c)/c$, as follows:

²⁰ Schulwas-Sorokin, E. D., & Posnov, M. V. Phys Rev 47: 166. 1935

Temperature range	c_i	t_i °C
50 to 32	136	25.3
32 to 24	178	23
-18 to -28	-93.8	-17.9
-28 to -42	-68.5	-20.6
-42 to -80	-47.9	-27.1
-80 to -140	-80.6	16.6

Below -160° C the dielectric constant is about 9 and is independent of temperature. Its largest value is reached at the two Curie points where values from 1400 to 10000 have been recorded.

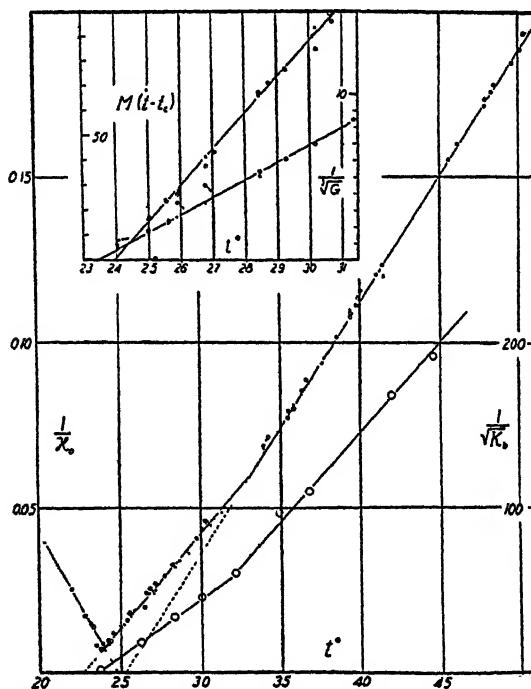


FIGURE 3. Variation of the reciprocal susceptibility and of the Kerr constant $K_0 - \frac{1}{2}$ of Rochelle salt with temperatures above the upper Curie point. Inset: Verification of $G - \frac{1}{2} = g(t - t_c)$ and of $(EX - \frac{1}{2})_{X=0} = M(t - t_c)$.

The older measurements in strong electric fields use either the ballistic galvanometer or the electrometer. Greater accuracy is obtained with the a. c. method of Sawyer and Towers¹⁴ in which the charge of the crystal condenser is recorded as function of the potential on the screen of a cathode ray oscilloscope. A series of photographs of these $P(E)$ curves are given in FIGURE 5. This figure illustrates

the gradual metamorphosis of the dielectric properties as the temperature is changed from the normal to the ferroelectric temperature range. At the lower Curie point the transition occurs in exactly the same manner. Above 32° C and below -26° C the crystal is "para-electric," *i. e.* the polarization is proportional to the applied field (up to 1500 Volt cm). With closer approach to the Curie points the initial slope of the $P(E)$ curves increases rapidly, it becomes infinite at these points, and simultaneously the curvature increases. At the Curie points the curves split to produce a hysteresis loop. Height and width of these loops increase very fast for temperatures near the Curie points, both are a maximum near 0°. For strong fields the

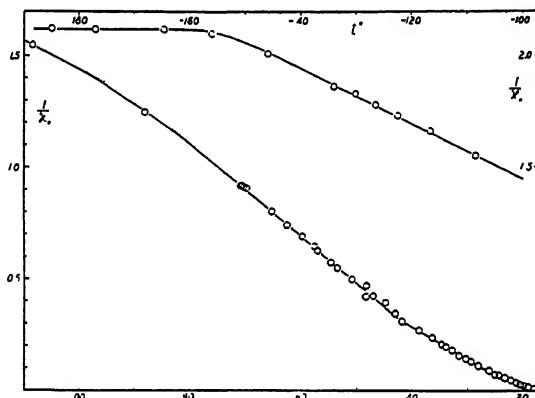


FIGURE 4. The reciprocal susceptibility of Rochelle salt below the lower Curie point.

polarization approaches saturation but it should be noted that even at 0° the "saturation curve" has a finite slope. Much higher fields would be required to effect complete saturation because at all temperatures ($\Theta - T$) is small.

A second method for dielectric investigations in strong fields has been developed by the writer.²¹ By measuring the capacity of the crystal condenser with a bridge while a constant D. C. voltage is applied across the electrodes one obtains the "reversible susceptibility" $\kappa_E = \partial P / \partial E$, *i. e.* the slope of the $P(E)$ curve for a series of field strengths E . A typical result is shown in FIGURE 6. The curve represents the derivative of a hysteresis loop. This second method has been used for a study of the dielectric properties in the transition

²¹ Mueller, H. Phys. Rev. 47: 175. 1935. This work was done in collaboration with Forbes, J. E. M.A. Thesis, M. I. T. 1934; Phys. Rev. 45: 736. 1934.

range from 32° to 24° . The results in FIGURE 7 show that for not too strong fields $\kappa_E = \kappa_0(1 - GE^2)$, where G increases rapidly with decreasing temperature, but this equation fails for stronger fields when the temperature is near the Curie point.

The most logical procedure for testing the phenomenological theory consists in comparing its consequences with the experimental results above the Curie point, where the dielectric properties are less complicated and where the data are more accurate. This comparison will furnish information about the numerical values of the parameters

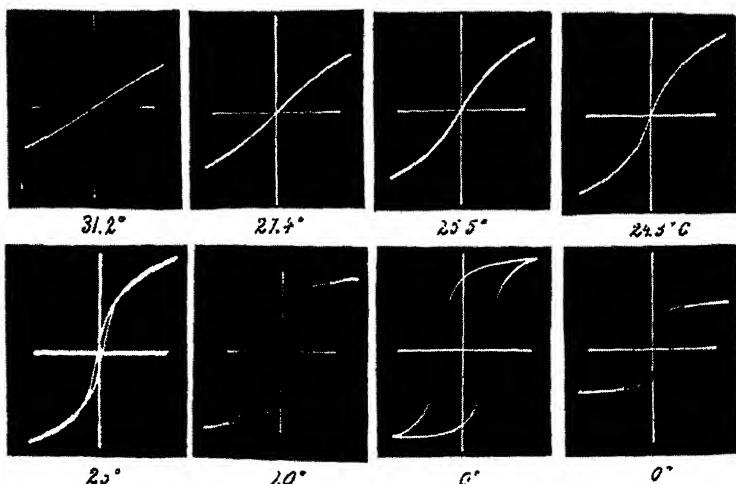


FIGURE 5. Polarization curves of Rochelle salt. Upper row for temperatures above the upper Curie point, lower row Hysteresis loops for temperatures between the Curie points.

f , Θ , and β . As a final test the theory should then be able to account for the behavior below the Curie point.

From the fact that near the Curie point $1/\kappa_0 = (t - t_c) / 178$ one infers that in this temperature range one can approximate $(T - \Theta) = \gamma(t - t_c)$ where $\gamma f = T_c$, $c = 1.67$. For not too large fields E we may assume that the solution of Equation (3) is given by a power series $P = \sum s_n E^n$. This method furnishes $\kappa_E = \kappa_0(1 - GE^2)$, in confirmation of the empirical relation. The theory requires that $G^{-1} = g(t - t_c)$, and $g = \frac{1}{2} (3\beta f)^{-1/2}$. The temperature dependence of G is verified in FIGURE 3 (inset) which gives $g = 1.05$. To explain the results for stronger fields it is convenient to introduce a new quantity $X = f^2(\kappa_0 - \kappa_E)/(1 + f\kappa_0)(1 + f\kappa_E)$. Since κ_0 and κ_E are

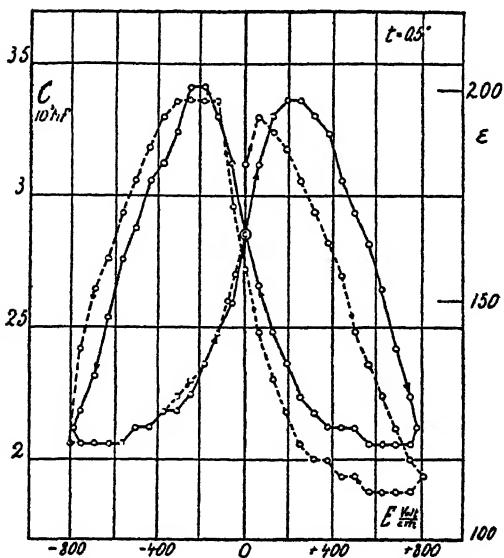


FIGURE 6. The reversible dielectric constant of Rochelle salt in the ferroelectric temperature range. The curve is the derivative of a hysteresis loop.

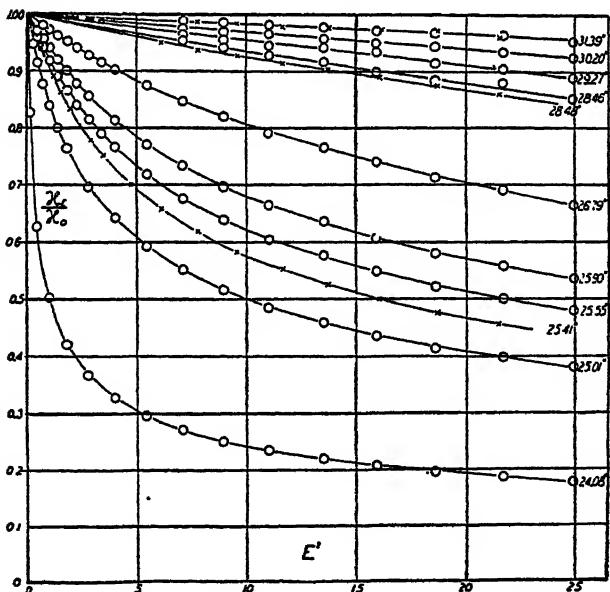


FIGURE 7. Variation of the reversible susceptibility of Rochelle salt with field intensity for temperatures above the upper Curie point.

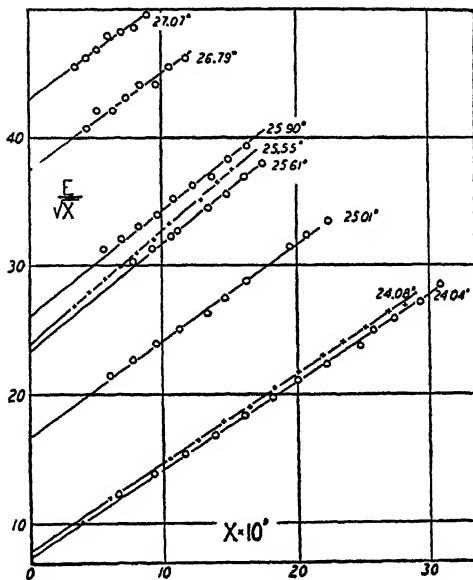


FIGURE 8. Verification of the law $EX - \frac{1}{2} = M(t - t_0) + NX$, whereby
 $X = (\epsilon_0 - \epsilon_X)/\alpha\epsilon_0$.

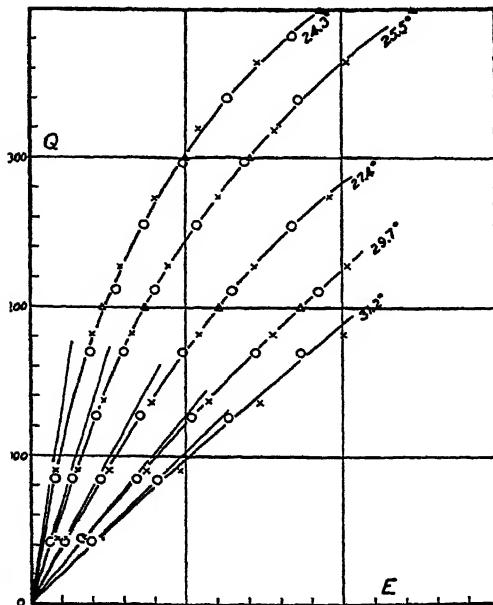


FIGURE 9. Theoretical polarization curves for Rochelle salt above the upper Curie point. O and X indicate experimental values calculated from the photographs in FIGURE 5.

large, X is for all practical purposes identical with $X = (\kappa_0 - \kappa_B)/\kappa_0 \kappa_B$. The theory predicts then $EX^{-\frac{1}{2}} = M(t - t_o) + NX$, $M = \gamma/fT_c(3\beta)^{\frac{1}{2}}$, $N = 1/3f^2(3\beta)^{\frac{1}{2}}$. The measurements verify also this relations (FIGURES 8 and 3 inset) and give $M = 13.5$, $N = 800$. The theory can be checked also by using the data furnished by the cathode ray oscilloscopes. FIGURE 9 shows that the relation between charge and potential as computed from the photographs in FIGURE 5 agrees with the theoretical curves.

Although these experiments furnish the values of 4 constants c , g , M , and N we cannot determine the 3 parameters, θ , f , β , because, in

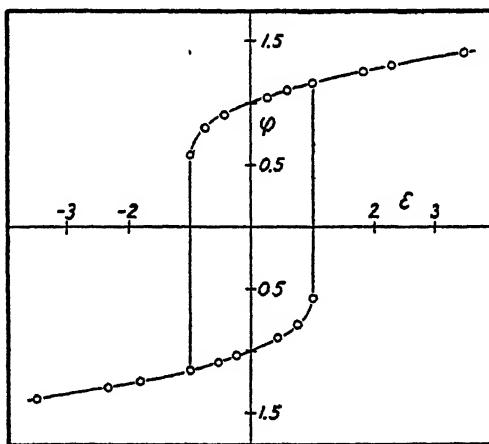


FIGURE 10. Theoretical hysteresis curve.

accordance with theory, they satisfy the relations $g^3 = M^2/c$ and $N = \frac{1}{3} cM$. The measurements furnish therefore only two quantities $\beta_f = 1.67$ and $\beta f^4 = 1/9cMN = 5.8 \cdot 10^{-8}$.

If the dipole theory is applicable and if one assumes that near the melting point (where $c = 136$) all dipoles can rotate freely we should have $f = T_c/c = 2.19$, $c = N\mu^2/3k$ and $\beta = N\mu^4/45(kT)^3$. The number of water molecules in 1 cc of Rochelle salt is $N = 1.52 \cdot 10^{22}$ and their moment (in the vapor) is $\mu = 1.85 \cdot 10^{-18}$. Hence on the basis of this theory $c = 127$, $\beta = 6 \cdot 10^{-11}$. Although the observed value of β is considerably larger it does not appear impossible to reconcile the results with this theory.

For the ferroelectric temperature range the phenomenological theory predicts a hysteresis loop of the form shown in FIGURE

10. This diagram gives $\varphi = F/F_0$ versus $\epsilon = E/E_c$, where $E_c = \frac{2}{3T(3\beta f T)^{1/2}} (\Theta - T)^{3/2} = \frac{2}{3} g^{3/2} (t_c - t)^{3/2}$ is the coercive force and $F_0 = fP_0$, where P_0 is the remanent polarization and should have the value $P_0^2 = (\Theta - T)/f^3 \beta T = \gamma f(t_c - t)/T f^4 \beta$. To calculate the curve for any particular temperature it is only necessary to adjust the scale factors according to the above equations. If one assumes that the

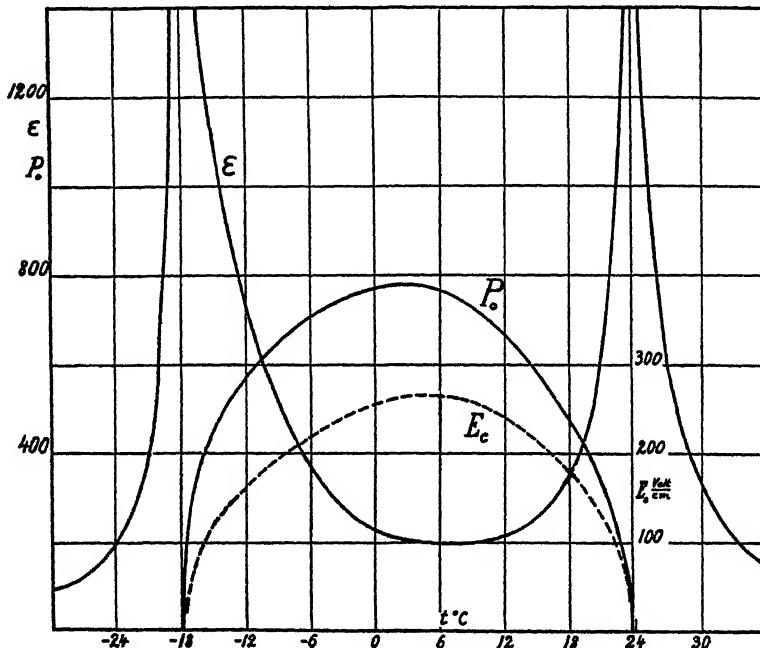


FIGURE 11. Remanent polarization P_r , coercive force E_c and dielectric constant for small fields of Rochelle salt in the ferroelectric temperature range.

bridge measurements record the slopes of the $P(E)$ loops at $E = 0$ the theory furnishes $\kappa_0 = \frac{T}{2f(\Theta - T)} - 1/f$ or approximately $1/\kappa_0 = 2(t_c - t)/c$, and the relation $P_0^2 \kappa_0 = 1/2f^4 \beta = 8.6 \cdot 10^6$ should hold.

The observed values²² of $\epsilon_0 = 1 + 4\pi\kappa_0$, P_0 and E_c are recorded in FIGURE 11. We note that the relation $P_0^2 \kappa_0 = 8.6 \cdot 10^6$ is surprisingly well satisfied. Similarly the modified Curie-Weiss law

²² From measurements by Bradford, E. B.S. Thesis, M. I. T. 1934. The remanent polarization P_r is obtained from cathode ray oscilloscopes by extrapolating the saturation curve to intersect with the P axis. (See FIGURE 5.)

seems to hold for temperatures near the Curie point. FIGURE 3 shows indeed that the slope of the $1/\kappa_0$ curve is twice as large below than above the Curie point. The course of the $P_0(t)$ and $E_c(t)$ curves, particularly their steep gradient near both Curie points, also correspond to the theoretical predictions. A closer inspection of the data reveals, however, that quantitatively the agreement is not

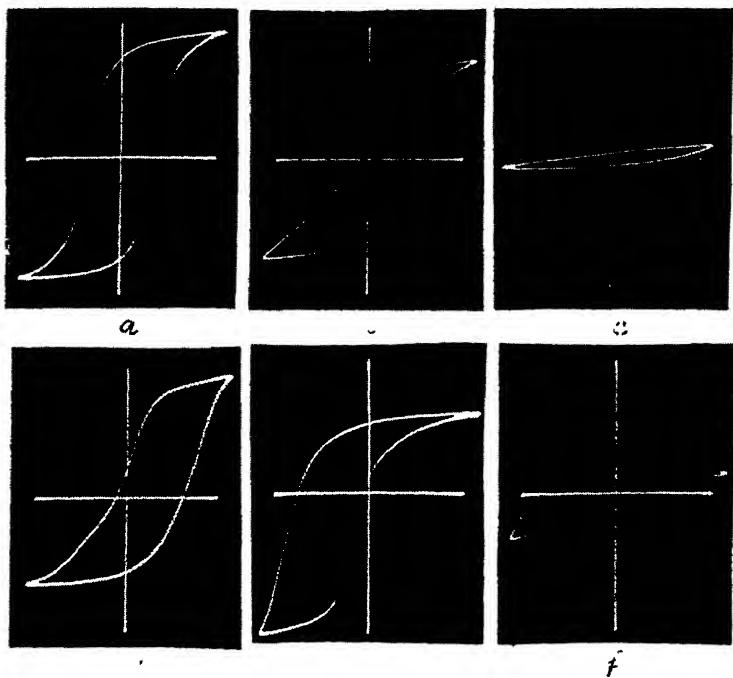


FIGURE 12 Hysteresis curves of strained crystals of Rochelle salt a free crystal, b. pressure prevents deformation, c same with larger pressure d pressure acts during half cycle, e same during other half cycle, f same with larger pressure Ten perature 0°, Max. field 2 k v/cm

satisfactory. In comparison with the theoretical values ($P_0 = 310$ stat U., $E_c = 200$ volt cm for $t_c - t = 1$) the measured values of E_c are about 10 times too small, those for P_0 are twice too small and κ_0 is found too large. These discrepancies may in part be due to the inadequacy of Equation (3). The observed P_0 values indicate that at 0° the inner field $F = fP_0$ is at least $1/2$ million volt cm, and it is therefore doubtful whether higher powers of F can be neglected. On the other hand we should consider also the possibility that the data are inaccurate or that the interpretation of the data is inadequate.

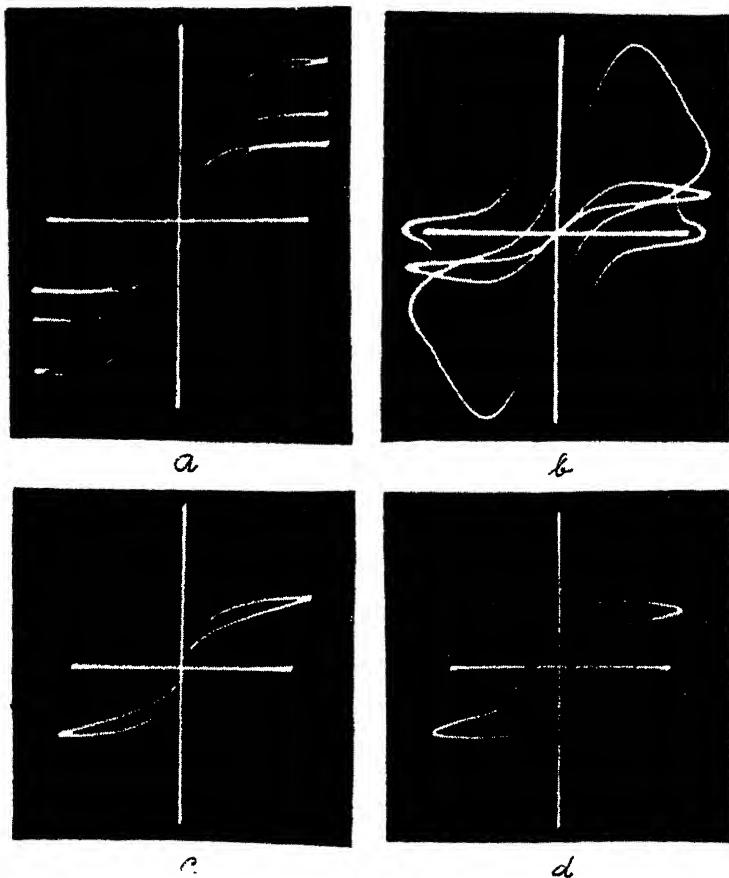


FIGURE 13. Influence of circuit constants on the hysteresis loop oscillograms: a. variation of measuring condenser, b. variation of compensating resistor, c. same as b., d. this loop is transversed in the "wrong" direction.

It is obvious that the theoretical curve (FIGURE 10), which has an infinite slope, cannot be verified by experiments. The steepest slope of the observed hysteresis curves corresponds to a dielectric constant of between 30000 to 100000.²³ The form and shape of the curves depend somewhat on the dimensions of the crystal and the frequency of the field and they are very sensitive to small strains (FIGURE 12). The conductivity of the crystal influences the width of the loops. It is impossible to correct properly for this error by introducing a

²³ David, E. Helv. Phys. Acta, 8: 431 1935.

compensating resistance into the measuring circuit, because the capacitance of the crystal condenser varies between large values (ratio 1 : 5000) during each cycle. By varying the circuit constants the loops can be deformed to very grotesque figures (FIGURE 13). It is therefore quite likely that the measured E_0 values are not reliable. The measurements of P_0 and of κ_0 are not affected by this difficulty. That they are fairly accurate is shown by the fact that the data of the various observers in America, Russia and Switzerland differ by less than 20%.

The difference between the theoretical and observed values of P_0 can probably be ascribed to a fundamental difference between the theoretical and experimental concepts of polarization. The measurements record as polarization the induced surface charges on the electrodes. The theory, however, considers the volume polarization within the crystal. In normal dielectrics and in Rochelle salt above the Curie point these two concepts are identical. In the ferroelectric state, however, they may differ. This fact is well known in the field of ferromagnetism where every Weiss region is polarized, though the substance as a whole may not show any magnetization. Rochelle salt probably also has its Weiss regions, some are polarized in the $+a$ and others in the $-a$ direction and the measurements record only the algebraic sum of the polarizations of all regions. If one takes into account that there must be a strong interaction between neighboring regions the discrepancies between theory and observation are not altogether incomprehensible, although a further clarification is desirable.

THE PYROELECTRIC EFFECT

According to the dielectric data the polarization of Rochelle salt in the ferroelectric temperature range does not vanish when the field is removed. The crystals should therefore show a spontaneous polarization equal to the remanent moment. To be sure, if both kinds of Weiss regions were present in equal numbers in all crystals, this polarization could not be observed. The facts indicate that this is not the case in small crystals, though it may occur in large samples with areas larger than 2×2 cm. Due to the spontaneous polarization opposite faces normal to the a axis become oppositely electrically charged. Under normal condition these charges cannot be observed because compensating charges are carried along the surface. If, however, the crystal is cooled rapidly from 25° to 0° the polarization increases so fast that the small conductivity cannot produce charge

compensation and appreciable potential differences are created. Sparks of a few millimeter length can thus be produced between wires attached to the electrodes. If powders with two kinds of oppositely charged and differently colored particles are sprayed on the crystal, the two faces normal to the a axis become differently colored. In the case of large crystals the colors frequently form a regular checker board on one surface. In small crystals the spontaneous polarization always has the same direction, even if the crystal is cooled in an opposing electric field or if it is heated to near its melting point between successive tests.

To measure this pyroelectric effect one measures the charges which are developed during a very slow and small temperature change. This is done either by counting how often they can charge a large measuring condenser to a small fixed potential (1/10 volt, Gaugain's method) or by measuring the compensating charge required to prevent the creation of a potential difference across the electrodes. The compensation method is more accurate and has lately been improved by using a vacuum tube, instead of an electrometer, for the potential control.²⁴

FIGURE 14 presents the results for two crystals and FIGURES 27 and 28 contain additional data obtained with another sample. Contrary to expectation, most small crystals give the same results, within an error limit of about 20%, and the pyroelectric moment is very nearly the same as the remanent polarization as measured from the hysteresis loop of much larger crystals. Near the upper Curie point the theoretical law $P_0^2 = h(t_c - t)$ is satisfied (see FIGURE 14 inset), but the observed value $h = 2.3 \cdot 10^4$ (according to the latest data) is four times smaller than the theoretical value $\gamma f/T_0 \beta f^4 = 9MN = 9.7 \cdot 10^4$. The difference can again be blamed on the oppositely polarized Weiss regions, but it is difficult to understand, on this basis, why the results are so consistent. It is only in large crystals where the pyroelectric effect becomes very erratic, but this can be due to the difficulty of uniform heating.

The spontaneous polarization assumes tremendous values. Fields of the order of magnitude of several million volts/cm would be required to create such large polarizations in ordinary solids. According to the measurements of Busch¹⁶ the crystals of the second ferro-

²⁴ The principle of the method is due to Ackermann, W. Ann. d. Physik. 46: 197. 1915. Measurements on Rochelle salt, using an Edelmann electrometer, were made in collaboration with Tarnopol, L. M.S. Thesis, M. I. T. 1934. The tube method, employing a Western Electric D-96475 electrometer tube, was developed by Saunders, H. O. M.S. Thesis, M. I. T. 1939.

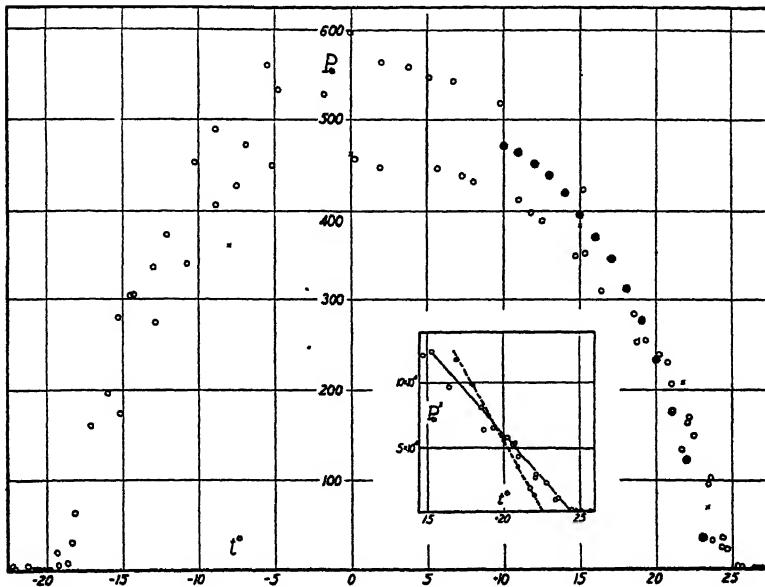


FIGURE 14. Pyroelectric polarization of Rochelle salt. \circ measurements with electrometer, \bullet with electrometer tube.

electric group become between 50 to 100 times stronger polarized than Rochelle salt.

THE PIEZOELECTRIC EFFECT

Due to its importance for technical applications the piezoelectric effect of Rochelle salt has been the object of a large number of investigations. They include the direct effect, where the polarization is measured as a function of applied strains or stresses, and the inverse effect, where the deformation of the crystal is studied as function of the electric field. The observations verify, in a general way, the assumption (4) that the strains are proportional to the inner field $F = E + fP$. Since in the investigated temperature range fP is always much larger than E , the strains for small fields will be proportional to the polarization, and we can write

$$y_z = d_{14}fP = \delta_{14}P = \delta_{14}\kappa_0 E.$$

where δ_{14} is the piezoelectric strain coefficient.²⁵ Hence we expect

²⁵ According to its definition δ_{14} differs from the piezoelectric constants of the classical theory. Recently Mason, W. P., Phys. Rev. 55: 775. 1939, has given a new verification of the proportionality between P and y_z . Mason gives $-Y_z = f_{14}P$, $f_{14} = 7.8 \cdot 10^4$. Hence, since $-Y_z = c_{44}/s_z$, where c_{44} is an elastic constant, we should have $\delta_{14} = f_{14}/c_{44}$. From Mason's data we derive $\delta_{14} = 62 \cdot 10^{-6}$ which is of the same order of magnitude but 3 times larger than the value derived from Norgorden's data.

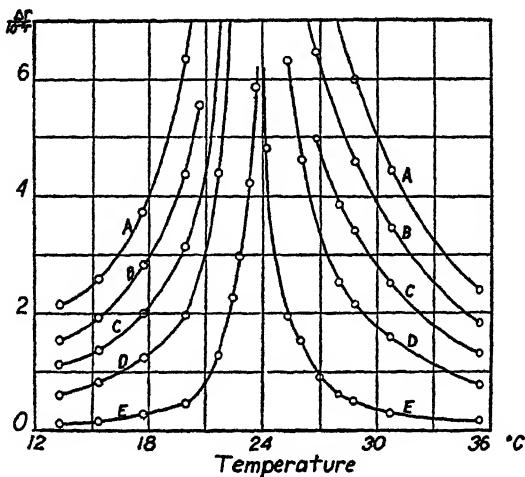


FIGURE 15. Inverse piezoelectric effect in Rochelle salt for various field intensities: A : 47.2, B : 37.7, C : 18.9, D : 6.3 volts/cm. Frequency 700 cycles. (acc. to Norgorden).

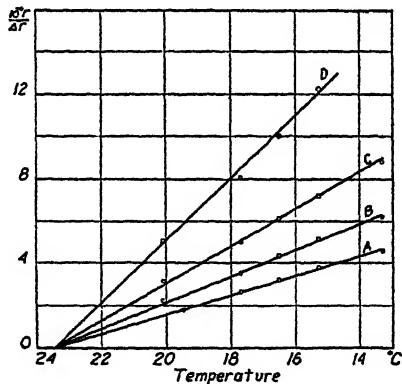


FIGURE 16. Verification of the Curie-Weiss law for the inverse piezoelectric effect in Rochelle salt for temperatures above the Curie point. (acc. to Norgorden).

that the temperature variation of y_s is the same as that of the susceptibility κ_0 . The measurements of Norgorden²⁶ verify this conclusion. FIGURE 15 gives $\Delta r/r = \frac{1}{2}y_s$ as function of temperature for various field strengths. The reciprocal $r/\Delta r$ is expected to satisfy a Curie-Weiss law $r/\Delta r = \frac{k}{E}(t - t_c)$ at temperatures above the Curie point, and $r/\Delta r = 2k(t_c - t)/E$ in a short range

²⁶ Norgorden, O. Phys. Rev. 49: 820. 1936.

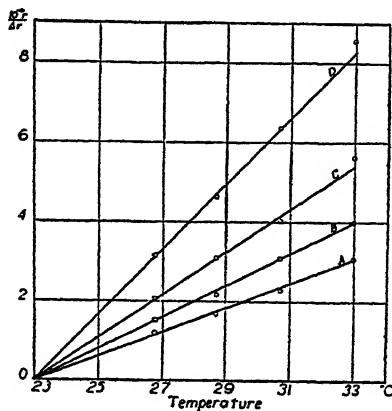


FIGURE 17. Verification of the modified Curie-Weiss law for the inverse piezoelectric effect in Rochelle salt for temperatures below the upper Curie point. (acc. to Norgorden).

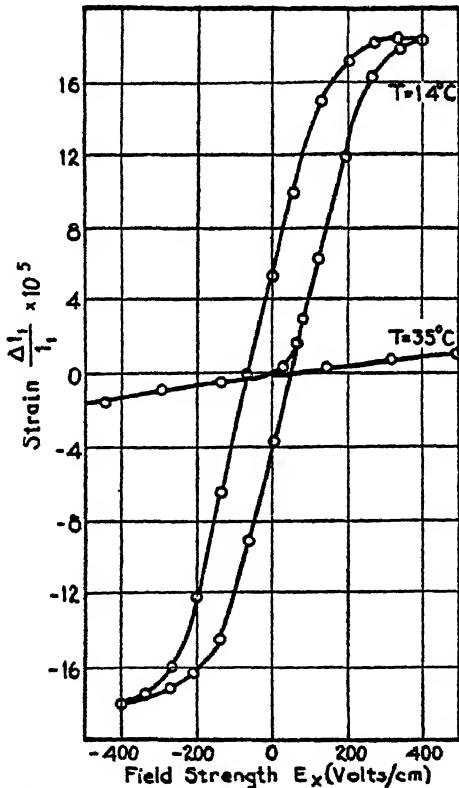


FIGURE 18. Hysteresis loop of the inverse piezoelectric effect in a mixed Rochelle salt crystal containing 0.37 per cent Na-Tl tartrate. (acc. to Bloomenthal)

below this point. Within the probable errors this is verified, as shown in FIGURES 16 and 17. These data furnish $\delta_{14} = 22 \cdot 10^{-8}$.

In strong fields the $y_s(E)$ curves are similar to the $P(E)$ curves. At temperatures between the Curie points the curves form hysteresis loops²⁷ (see FIGURE 18). No sufficient data for a quantitative test of the theory are available.

In the ferroelectric range the spontaneous polarization P_0 is expected to create a spontaneous strain $\delta_{14}P_0$. This spontaneous deformation has not as yet been observed. It should manifest itself as a change of the angle between the crystallographic b and c axes and should amount to at least $30''$ at 0° .

OTHER LINEAR EFFECTS

It is to be expected that other effects, which in normal substances are proportional to the electric field, will in Rochelle salt show a behavior similar to the piezoelectric effect. A study of the linear electro-optical effect, which has not been investigated since Pockels' discovery, would therefore be of great interest. Other linear effects, which are not directly related to the electric properties, as *f. i.* elasticity, photoelasticity and temperature expansion have no large anomalies in Rochelle salt, though Davies²⁸ reported a sudden change of about 1% of the elastic constants occurring at the upper Curie point. It is probable that the constants of all these effects undergo slight changes when the crystal is in an electric field or when it becomes spontaneously polarized at the Curie points, because x-ray investigations have shown that a field alters somewhat the crystal structure. Staub²⁹ and Nemet³⁰ have observed a change of the intensity of certain x-ray diffraction spots when an electric field is applied, and Staub³¹ reports a similar change due to the spontaneous field (see FIGURE 19). These intensity changes are reversed with a reversal of the field. They must be ascribed to displacements of atoms within the unit cell.

An interesting verification of the pyroelectric effect in Rochelle salt is furnished by the linear electro-caloric effect. It is well known that in a spontaneously polarized medium the sudden application of an electric field ΔE produces an adiabatic temperature change ΔT

²⁷ Bloomenthal, S. Physics 4: 172. 1933.

²⁸ Davies, R. M. Nature 120: 332. 1927.

²⁹ Staub, H. Helv. Phys. Acta. 7: 1934; Phys. Zeits. 34: 292. 1933.

³⁰ Nemet, A. Helv. Phys. Acta. 8: 1935.

³¹ Staub, H. Phys. Zeit. 35: 720. 1934.

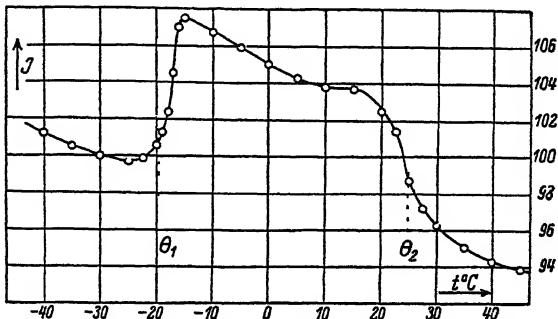


FIGURE 19. Variation of the intensity of the (222) x-ray reflection in Rochelle salt with changes of temperature in the ferroelectric range. (acc. to Staub).

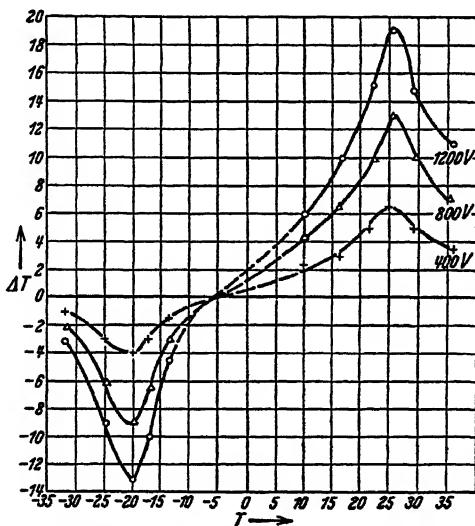


FIGURE 20. Adiabatic change of temperature with application of an electric field, caused by the electro-caloric effect in Rochelle salt. (acc. to Kobeko and Kurtschatow).

$= -\frac{T}{c} \left(\frac{\partial P_0}{\partial T} \right)_E \Delta E$, where c is the heat capacity of 1 cc. The existence of this effect in Rochelle salt has been established by Kobeko and Kurtschatow.³² In accordance with the observed $P_0(T)$ curve they find (FIGURE 20) that the effect is largest at the Curie points and reverses its sign near 0° . The order of magnitude of the temperature changes (10^{-5} °C for a field of 1 volt/cm) agrees with the formula but the data are not complete enough for a quantitative comparison.

³² Kobeko, P. P., & Kurtschatov, J. Zeit. Physik. 68: 192. 1930.

THE KERR EFFECT

A new type of anomalies will occur in Rochelle salt for effects which normally are proportional to the square of the field strength. A study of the Kerr effect offers therefore an independent test of the phenomenological theory. This effect is measured by sending a beam of monochromatic light through the crystal. The incident light is polarized at 45° to the a axis, it transverses the crystal parallel to either the c or b axis and the state of polarization of the transmitted beam is determined with a Babinet compensator. This instrument serves to measure the changes Δ_b and Δ_c of the birefringence $(n_a - n_b)/\lambda$ or $(n_a - n_c)/\lambda$ which arise when a field is applied in direction of the a axis. n_a , n_b , n_c are the refractive indices for light vibrating parallel to the a , b or c axis, and λ is the wavelength of the light.

FIGURES 21 and 22 show the dependence³³ of Δ_b and Δ_c on the field at various temperatures above the upper Curie point. For both light directions the results are similar but Δ_b is 40% larger than the corresponding value of Δ_c . Hence the "Kerr constants" ρ in Equation (5) have the ratio $\rho_b/\rho_c = 1.4$ and we conclude that $\Delta_b - \Delta_c = \delta \left(\frac{n_b - n_c}{\lambda} \right) = \Delta_a$ is a similar function of E and t . There exists therefore in Rochelle salt a "longitudinal" Kerr effect, i. e. a change of birefringence for light passing parallel to the electric field. Its existence has been verified directly by sending a light beam through holes in the electrodes.

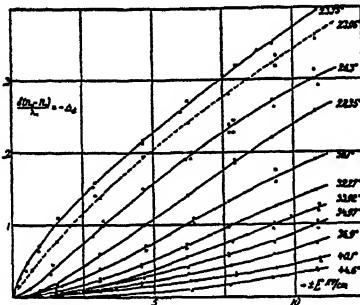
For small fields and high temperatures the Kerr effect is proportional to E^2 , for stronger fields the curves assume S shape and at the Curie point the simple law $\Delta = D E^{2/3}$ is satisfied. This is exactly the behavior predicted by the theory. From Equations (3) and (5) it follows that for small fields $\Delta = \rho T^2 E^2 / (T - \Theta)^2 = KE^2$ and at the Curie point $\Delta = \rho \left(\frac{E}{\beta f} \right)^{2/3}$. According to these equations

$1/\sqrt{K} = \frac{1}{f\rho^{1/2}} \frac{1}{\kappa_0}$ must have the same temperature dependence as $1/\kappa_0$. This is verified in FIGURE 3 which furnishes an average value of $f^2 \rho_b = 6 \cdot 10^{-7}$ (it varies from $5.1 \cdot 10^{-7}$ at high temperatures to $6.6 \cdot 10^{-7}$ near the Curie point³⁴). From the measurements at the Curie

³³ The electro-optical and optical measurements were made in collaboration with Groat, L. M. M.S. Thesis, M. I. T. 1932.

³⁴ Whence $\rho_b f^2 = (4.3 \pm 0.5) \cdot 10^{-7}$. This value was given incorrectly in a previous publication: Phys. Rev. 47: 189. 1935.

point we deduce $D = \frac{\rho_b}{(\beta f)^{2/3}} = 3.5 \cdot 10^{-2}$. From these two numerical values we find $\beta f^4 = (6.9 \pm 1.5) \cdot 10^{-8}$. Within the large probable error this result agrees with that derived from the dielectric data which gave $\beta f^4 = (5.8 \pm 0.7) \cdot 10^{-8}$. This agreement between two entirely independent sets of data proves the validity of the phenomenological theory. The electro-optical data cannot furnish any new



tion, which was first suggested to the writer by Professor Debye, is as follows: Since the crystal has a spontaneous inner field $F_0 = fP_0$ the Kerr effect is $\Delta = \rho(F_0 + E)^2 = \rho F_0^2 + 2\rho F_0 E + \rho E^2$. Since at a given temperature the first term is constant, and the last term is negligible ($F_0 >> E$), the measurements record only the term $2\rho F_0 E$ which is linear in E and does not reverse its sign, because in

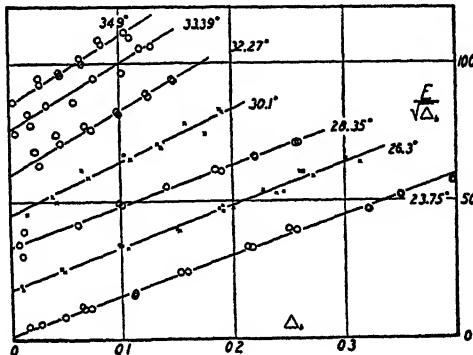


FIGURE 23. Verification of the law $E\Delta - \frac{1}{2} = R(t - t_c) + Q\Delta$ for the Kerr effect of Rochelle salt above the upper Curie point.

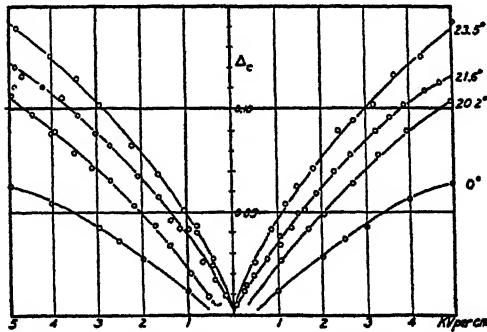


FIGURE 24. Kerr effect in Rochelle salt for temperatures in the ferroelectric range.

a strong field the inner field F_0 reverses its sign with a reversal of E . A more careful study of the data reveals, however, that this argument is not adequate. The slope of the $\Delta(E)$ curves, instead of increasing, actually decreases with increasing field strengths. Furthermore, the slope is smaller at 0° than at 20° C while the above derivation predicts the largest slope to occur at 0° where F_0 is a maximum. The correct explanation is found by using the fundamental equations (3) and (5). For any quadratic effect they pre-

dict the behavior illustrated in FIGURE 25. For small fields we must find a quadratic hysteresis curve. It is difficult³⁵ to measure this loop with any degree of accuracy but the results in FIGURE 26 demonstrate its existence. The strong fields used in electro-optical measurements correspond to fields in the "saturation" branches of

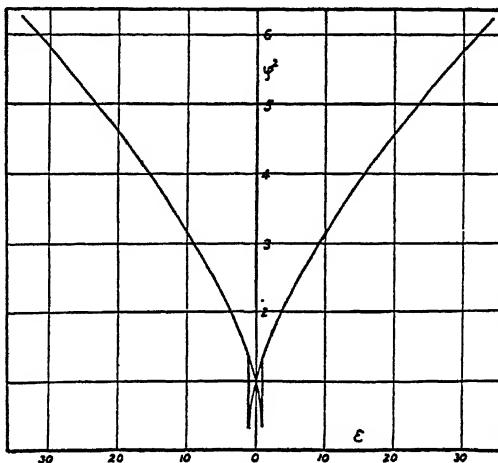


FIGURE 25. Theoretical hysteresis loop and "saturation" curve for quadratic electric effects in Rochelle salt.

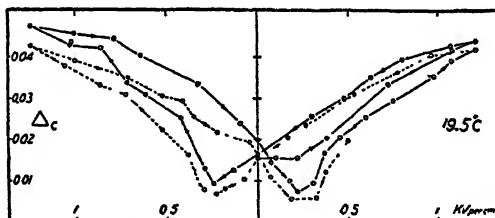


FIGURE 26. Quadratic hysteresis loop of the Kerr effect in Rochelle salt.

the hysteresis loops, but saturation is never reached because the temperatures are too near the Curie temperature. The closest approach to saturation occurs when the temperature is near 0°, or when a strong field is applied. This explains the observed fact that under these conditions the slope of the $\Delta(E)$ curves is reduced. The theory predicts for the initial slope of the $\Delta(E)$ curves $\left(\frac{d\Delta}{dE}\right)_{E=0} =$

³⁵ The difficulty arises from the fact that in these static measurements the precaution of applying the field for a very short time cannot be followed because observations of the hysteresis loops require a continuous change of the field.

$\sqrt{\frac{L}{(t_c - t)}}$ where $L = \rho^2 T_c / f \beta \gamma = 1/RQ$. The data give $L_c = 3.2 \times 10^{-4}$. Since $R_c = 10$, $Q_c = 310$ the theory is therefore verified.

THE SPONTANEOUS KERR EFFECT

In the ferroelectric temperature range the spontaneous polarization gives rise to a spontaneous Kerr effect $\Delta_0 = \rho f^2 P_0^2$. This effect is superposed upon the normal variation of birefringence with temperature and manifests itself in abrupt changes of the temperature gradient of the double refraction at the two Curie points. Δ_0 is found by subtracting from the observed values of $(n_a - n_b)/\lambda$

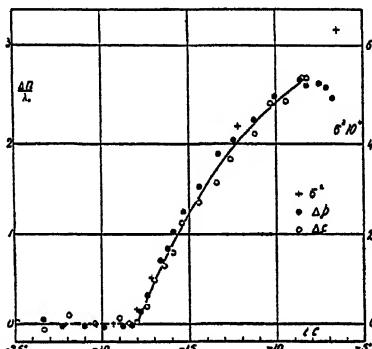


FIGURE 27. The spontaneous Kerr effect and the spontaneous polarization $\sigma = P$ (Gauvin's method) at the lower Curie point of Rochelle salt.

and $(n_a - n_b)/\lambda$ the values corresponding to the normal change, which are obtained by extrapolation from the measurements above the upper and below the lower Curie points. The results in FIGURES 27 and 28 show how sharply this effect sets in at the Curie points and how accurately Δ_0 can be measured. They demonstrate the expected proportionality between Δ_0 and the observed value of P_0^2 . According to the theory the factor of proportionality should have the value $\rho_b f^2 = 6 \times 10^{-7}$; according to the data this factor is at least 3 to 5 times larger. This discrepancy is not surprising. We have pointed out that the observations furnish too small values for the polarization P_0 because the induced surface charge depends on the number of Weiss regions and the orientation of the volume polarization within these regions. The spontaneous Kerr effect, however, is not influenced by these factors because it is a quadratic effect. Since, furthermore, the Kerr effect is determined by the conditions

in the interior of the crystal, it provides a much more reliable method for obtaining information about the volume polarization within Rochelle salt. If, therefore, we are able to show that the spontaneous Kerr effect leads to a spontaneous polarization identical with that predicted by the theory, we gain not only a new and better confirmation of the theory, but we also justify thereby our discussion concerning the difference between volume and surface polarization.

For temperatures near the upper Curie point the theory predicts $\Delta_0 = \frac{\rho f^2}{\beta f^3 T} (\Theta - T) = \frac{R_b}{Q_b} (t_c - t) = 4.6 \cdot 10^{-2} (t_c - t)$. This agrees indeed very closely with the observed value $\Delta_0 = 0.04 (t_c - t)$. Or,

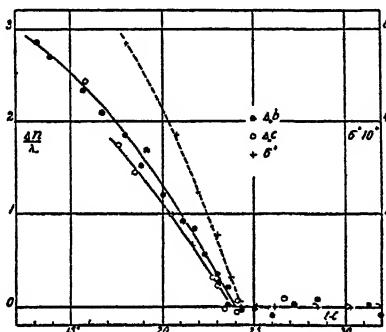


FIGURE 28. The spontaneous Kerr effect and the spontaneous polarization $\sigma = P_0$ at the upper Curie point of Rochelle salt.

vice versa, the optical measurements furnish for the spontaneous polarization near the Curie point $P_0^2 = \Delta_0/\rho f^2 = (7 \pm 1.5) \cdot 10^4 (t_c - t)$ which is only slightly smaller than the theoretical value $P_0^2 = \frac{\gamma f}{\beta f^4 T} (t_c - t) = (9.8 \pm 1.3) \cdot 10^4 (t_c - t)$ as deduced from the dielectric measurements above the Curie point.

THE SPECIFIC HEAT OF ROCHELLE SALT

Though we possess an unusually large amount of dielectric, piezoelectric, pyroelectric, optical and electro-optical data on Rochelle salt, they do not furnish enough information for the evaluation of the important Lorentz factor f . To determine its value we must resort to caloric measurements. In the ferroelectric state the heat capacity of a crystal differs from its normal value because energy is required to alter the spontaneous polarization. The additional heat capacity is

$$\Delta C = -\frac{1}{2}f \frac{\partial}{\partial t}(P_0^2)$$

The thermodynamical derivation of this relation is based on the assumption that there exists a functional relationship between P and $(E + fP)$. The equation must therefore be valid for Rochelle salt.

It should, perhaps, contain an additional term $-\frac{1}{2} \frac{\partial}{\partial t}(c_{44}y_s^2)$, which arises from the fact that an additional energy is needed to change the spontaneous deformation. However this correction is negligible because $c_{44}y_s^2 = c_{44}\delta_{14}^2 P_0^2 = 6 \cdot 10^{-3} P_0^2$ amounts to less than 1% of fP_0^2 . This effect should manifest itself in a sudden increase of the specific heat at the upper Curie point and a similar decrease at the lower Curie point. On the basis of our previous data this shift should be less than $2.3 \cdot 10^{-3}$ joules/gr = 0.15 f cal/Mol at the upper Curie point. A slightly larger value is found for the decrease at the lower Curie point.

The change of the specific heat at the upper Curie point was first predicted and measured to 5 cal/Mol by Kobeko and Nelidow³⁶. Later Rusterholz³⁷ confirmed this result. A value of $\Delta C = 5$ cal/Mol would lead to $f > 30$ and seems altogether too high. Two other features of Kobeko and Nelidow's and of Rusterholz's results do not agree with the dielectric data. For the temperature of the Curie point they find 26.2° and 25.8° C respectively, while in all other measurements it is below 24° . Below the Curie point the observed values of ΔC vanish much faster with decreasing temperature than Equation (6) could account for on the basis of any of the measured P_0 curves. It is difficult to account for these discrepancies, though a number of causes might be suspected, e. g. impurities, pressure, condition $E = 0$ not satisfied, etc. To clarify the situation three new investigations of the specific heat of Rochelle salt were undertaken by Hicks and Hooley,³⁸ Wilson³⁹ and Wildberger.⁴⁰ Neither of these investigations could verify the large ΔC values, in fact neither found any discontinuity of the specific heat at the upper Curie point and only Wilson detected a small negative ΔC at the lower Curie point (see FIGURE 29). These results do not imply that the effect does not exist, but rather that it is too small to be measured

³⁶ Kobeko, P. P., & Nelidow, J. G. Physikal. Zeit. Sov. Union. 1: 382. 1932

³⁷ Rusterholz, A. A. Helv. Phys. Acta. 8: 39. 1934.

³⁸ Hicks, J. F. G., & Hooley, J. G. Jour. Am. Chem. Soc. 60: 2994. 1938

³⁹ Wilson, A. J. C. Phys. Rev. 54: 1103. 1938.

⁴⁰ Not published, quoted in Reference 11.

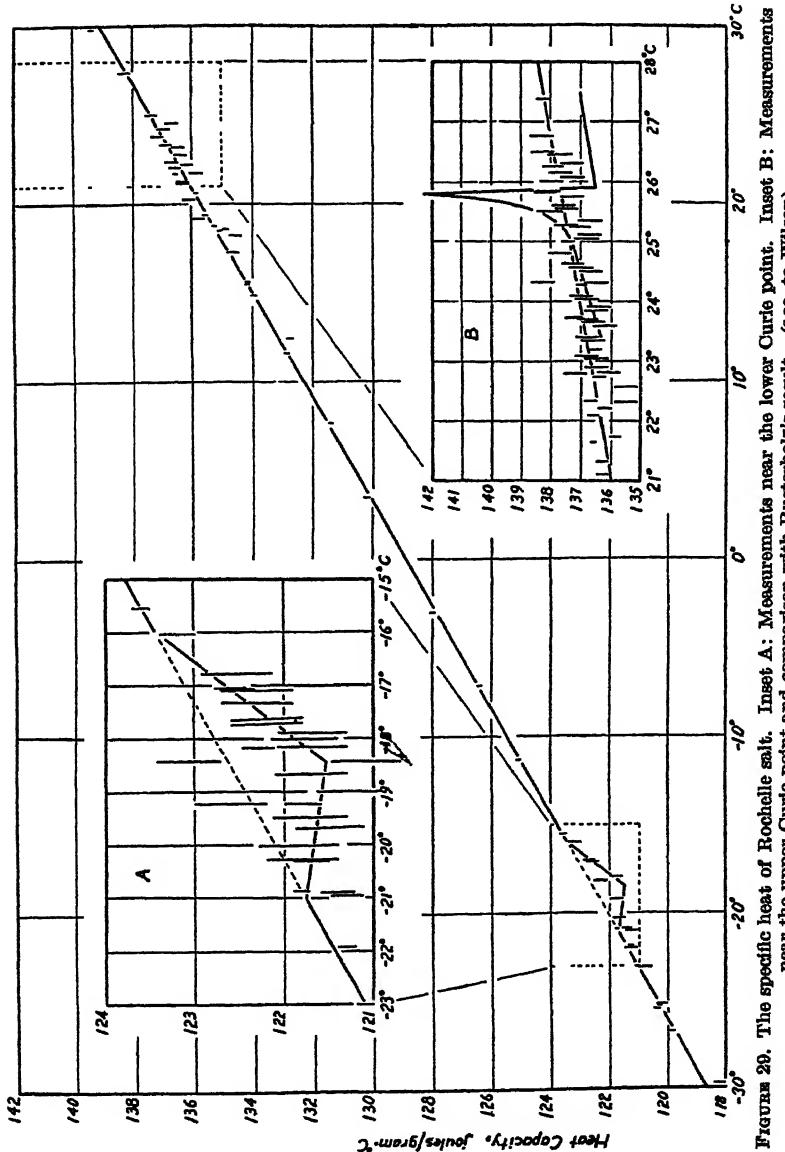


FIGURE 29. The specific heat of Rochelle salt. Inset A: Measurements near the lower Curie point. Inset B: Measurements near the upper Curie point and comparison with Rusterholz's result. (acc. to Wilson).

by the calorimetric methods used. Since these methods have a probable error of about 1%, it is concluded that the maximum value of ΔC at the upper Curie point is certainly less than 1 cal/Mol and probably less than $\frac{1}{2}$ cal/Mol. Consequently f cannot be appreci-

ably larger than $\frac{4}{3}\pi$ and the new measurement are compatible with all other data on Rochelle salt.

Stevenson and Hooley⁴¹ recently have measured a very large change of the specific heat at the Curie point of H₂KPO₄. Since this crystal has a much larger permanent polarization than Rochelle salt its large ΔC value can be accounted for without assuming an excessive f value; in fact the data lead to $f = 0.7$.

CONCLUSION

The experimental research of the properties of Rochelle salt has furnished a fairly consistent set of data which can be correlated by a simple phenomenological theory. A clearer understanding of the ferroelectric phenomena and of the empirical constants of the theory can only be obtained after the crystal structure of Rochelle salt has been determined. Since the substance is made up of many light atoms the prospects for an early solution of this problem are rather remote.

It appears more likely that further progress in this field can be made through a study of the second group of ferroelectric crystals H₂KPO₄, H₂KAsO₄ etc. Their lattice structure is known and is rather simple. They have a much larger polarization and it should therefore be easier to obtain accurate data. Hence the writer ventures the prediction that the problem of ferroelectricity will find its solution in a theoretical investigation of the binding forces and lattice properties of these crystals.⁴²

⁴¹ Stevenson, C. C., & Hooley, J. G. Phys. Rev. 56: 121. 1939. The same result has since been reported also by Bantle, W., & Scherrer, P. Nature 143: 980. 1939, and by Mendelsohn, J., & Mendelsohn, K. Nature 144: 595. 1939.

⁴² This paper presents the state of the problem of ferroelectricity at the time of the conference in the spring of 1939. Since then, partly as a result of the discussions during the conference, considerable progress toward a final solution have been made. Mason, W. P. Phys. Rev. 55: 775. 1939; 58: 744. 1940, and the writer Phys. Rev. 57: 829. 1940; 58: 565. 1940, have analysed the role of the piezoelectric effect and have shown that ferroelectricity is a true dielectric anomaly. The inner field theory has been abandoned in favor of a new interaction theory and the Kerr-effect is interpreted as a "morphic" effect. (Mueller, H. Phys. Rev. 58, Nov. 1. 1940). The structure of Rochelle salt has been determined by Beavers, C. A. & Hughes, W. Nature 146: 96. 1940. Slater, J. C. (to appear soon) has extended the statistical theory of the hydrogen bonds, which was presented by Dr. L. Onsager at the conference, and has given a new theory of the ferro-electric transition in H₂KPO₄.

ROTATION OF SOME LARGE ORGANIC MOLECULES¹

BY S. O. MORGAN

From the Bell Telephone Laboratories, New York, N. Y.

It has been observed experimentally that molecular rotation in the crystalline solid state, as shown by the existence of high dielectric constants, is a property of a large number of organic compounds of very different chemical structure. The earliest work on this subject appeared to indicate that this type of molecular rotation was to be found only among small compact molecules. This paper describes experimental results which indicate that the geometrical symmetry of the molecule rather than size alone is the most important factor in determining the ability of molecules to rotate in crystalline solids.

The materials which are considered belong to four different chemical classes and are all solids at room temperature, some of them having melting points above 200° C. All of the measurements of solids were made on pressed sheets having tinfoil electrodes affixed. The dielectric constant measurements on these solids are similar in many respects to the behavior usually observed in liquids. These measurements indicate that with respect to molecular interactions and the properties which are related to rotational motions of the molecules there is no sharp distinction between liquids and solids.

One of the prominent features of the dielectric behavior of polar liquids is anomalous dispersion. Anomalous dispersion may be described in terms of the static and infinite frequency dielectric constants and the time of relaxation of the polarization. If the relaxation time is short compared to the period of the measuring field the dipole or other absorptive polarization has time to build up before the field reverses and thus to contribute to the dielectric constant which measures the total dielectric polarization from all causes. This is what is called the static dielectric constant. However, if the relaxation

¹ The work described here is the result of a cooperative study of dielectric behavior which has been in progress at Bell Laboratories for several years. It is largely the work of Messrs. Addison H. White, W. A. Yager and B. S. Biggs. Some of the results have been described earlier in the following references: Yager, W. A., & Morgan, S. O. Jour. Am. Chem. Soc. 57: 2071-78. 1935. White, A. H., & Morgan, S. O. Jour. Am. Chem. Soc. 57: 2078-86. 1935. White, A. H., & Morgan, S. O. Jour. Chem. Phys. 5: 655-65. 1937. White, Addison H. Jour. Chem. Phys. 7: 58-60. 1939.

time of the dipole polarization is of the same order as or longer than the period of the applied field, the polarization will not have time to build up completely, or possibly even at all, and so there will be a smaller contribution, or none at all, to the dielectric constant arising from orientation of dipoles. The dielectric constant measured at frequencies too high for dipole or other absorptive polarizations to form is called the infinite frequency dielectric constant. The relaxation time is the ratio of the viscous or dissipative force constant of the force opposing the displacement of the dipole to the restoring force constant. If the viscous force is very low or the restoring force very high, the time of relaxation will be short. For liquids it has been customary to express the viscous force constant in terms of the coefficient of viscosity. For dipole polarizations the restoring force is the thermal energy tending to cause a return to random distribution.

Experimental studies show that anomalous dispersion is not limited to liquids or glasses but is also observed in crystalline solids in which there is no viscous flow. In describing the dielectric behavior of solids, then, it is preferable to discard the term "viscosity" and to describe the restriction of motion of the molecules or parts of molecules as due to potential barriers arising from the interactions between neighboring molecules. Thus the molecule, or the polar group, is restricted by barriers to a rotational oscillation and only when it acquires enough energy to overcome these barriers can it move to a new equilibrium position. This new position need not differ from the old by 180° , as in an end-for-end rotation, but may differ by any angle.

For each temperature there is a certain probability that the molecule or polar group will be able to acquire enough energy from thermal collisions to get over the barrier into a new equilibrium position. If this probability is high, then in a large number of the 10^{12} or so rotational vibrations per second which the molecule is normally undergoing it will succeed in surmounting the barrier. This would be a case of short relaxation time. If the energy of the barrier is very large compared to the thermal energy, the probability that the molecule or group will be able to surmount the barrier is small. This is a case of long relaxation time. Thus if the time of relaxation be short compared to the period of the measuring field, the molecule will appear to the field as if it were continually jumping from one equilibrium position to another and so there will be

a dipole contribution to the dielectric constant. If the time of relaxation be long compared to the period of the applied field, the dipole will contribute little or nothing to the dielectric constant. With this picture it is not necessary to have free rotation of the molecule or of a polar group but a restricted motion.

In some dielectrics there is evidence of a cooperative effect such that when one molecule starts to rotate it makes it easier for its neighbors to do so and the effect spreads. These materials are usually distinguished by a sharp transition above which rotation takes place with the same ease as in the liquid and does so up to very high frequencies. Below this transition there is no rotation even at very low frequencies. Many of the materials to be discussed show this cooperative effect; they have transitions, and at these transition temperatures there is evidence of a sharp change in density and heat capacity indicating a change in packing and freedom of motion of the molecules. Such density changes have not been observed in those solids which do not have transitions but change from a high to a low dielectric constant with anomalous dispersion. These latter materials apparently are not able to collapse to a more dense packing.

This picture is generally accepted for small compact molecules which are nearly spherical. One of the objects of this paper is to show that some very large molecules also are able to rotate, provided certain requirements are met as to symmetry.

The fact that the dielectric constant curve is practically continuous through the melting point in many cases is believed to be good evidence that it is the restricted orientation of the polar molecule which explains the high dielectric constant in the crystalline solid as well as in the liquid state. However, some other mechanism such as the rotational vibration of the polar group also may explain the result equally well in certain cases, but not in all.

The first figure shows the data for the dielectric constant plotted against temperature for d-camphor and a number of substituted camphors. These compounds differ chemically from camphor in having one or both of the hydrogens on the adjacent C atom substituted by polar groups. The melting point is marked by the heavy arrow. The curves given here have been somewhat smoothed out from the actual curves to eliminate some of the irregularities which complicate the picture in some minor respects but do not change it in its essentials. For example, there are in some cases hysteresis effects and dispersion effects, but only a single typical curve has been selected.

In two of these compounds, camphoric anhydride and camphor quinone, there is a considerable decrease of dielectric constant upon solidification but in the others there is no significant change. These compounds have different polar groups attached to the same carbon and all of them show a behavior similar to camphor but with different values of dielectric constant and different transition temperatures. The transition temperature is higher than that of camphor in every case. The temperature range in which the material is a solid but

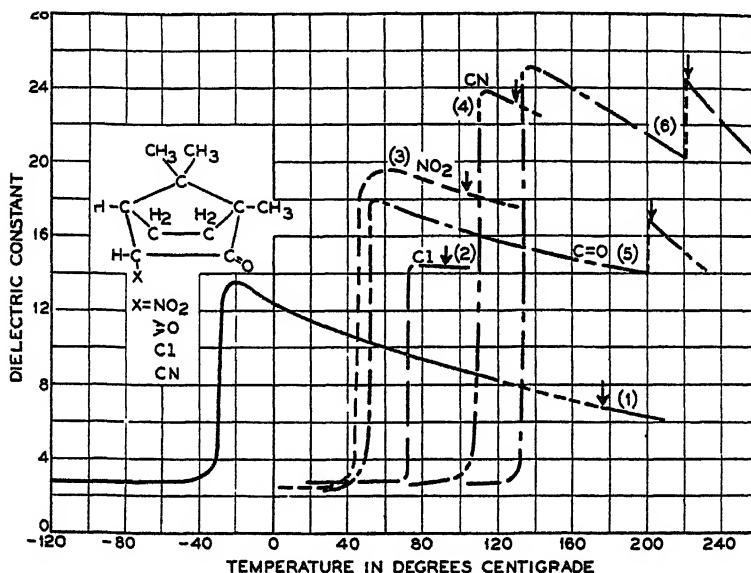


FIGURE 1. Curves of dielectric constant vs. temperature for camphor (1), chlorocamphor (2), nitrocamphor (3), cyanocamphor (4), camphor quinone (5) and camphoric anhydride (6). Melting point indicated by heavy arrows.

still has a high dielectric constant, and therefore presumably in which the molecules are still able to orient with much the same ease as in the liquid state, differs widely for these compounds. For chlorocamphor and also nitrocamphor it is about a 40° range but in camphoric anhydride, camphor quinone and camphor it is 85 , 150 and 215° , respectively. These are long enough ranges to make it certain that the effects are not due to the co-existence of liquid and solid phases.

Except for camphor, the electric moments of these compounds have not been measured but the calculated moments, using accepted values of group moments, are in the same order as the dielectric con-

stants and they are all higher than that of camphor. This is a good reason for believing that it is the rotation of the entire molecule rather than individual group rotations with which we are dealing.

FIGURE 2 shows the dielectric data for a number of other compounds related to d-camphor; in this comparison other groups have been substituted for the carbonyl oxygen. The curve for camphor has been repeated for reference to the data of FIGURE 1. Bornyl chloride has H and Cl groups instead of the C=O group. The dielectric constant is much lower than camphor, as it should be because of its lower electric moment, but it is higher than the square of the refrac-

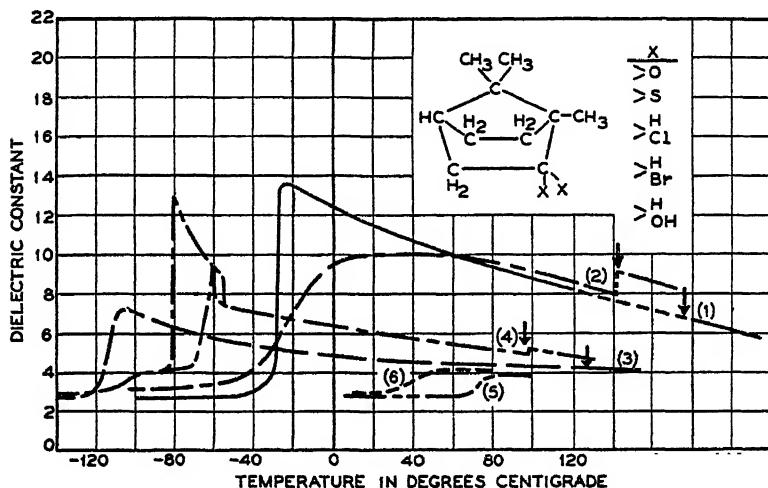


FIGURE 2. Dielectric behavior of camphor (1), thiocamphor (2), bornyl chloride (3), bornyl bromide (4), borneol (5) and isoborneol (6).

tive index and it retains its high value down to -115° , 245° below the melting point. Bornyl bromide shows an unusual behavior. The dielectric constant increases regularly in the solid state down to -55° , when it takes a jump to another form and then rises at a much more rapid rate down to -80° .

Borneol has H and OH groups instead of the carbonyl oxygen and shows a much lower value of dielectric constant although its electric moment, as measured in dilute solutions, is much the same as for bornyl chloride or bromide. Isoborneol, which differs in the exchange of positions of the CH₃ groups of the bridging carbon with two hydrogens of one of the ring carbons, likewise has a low value of dielectric constant. The data for the borneols suggested that perhaps

here we were dealing with only the rotation of the OH group and if the measurements were carried up to the melting point, a much higher value of dielectric constant would be found in the liquid. This was not the case. The value was lower, but such an experiment is unsatisfactory because of the high melting point of borneol and the fact that the dielectric constant of any polar material normally decreases with increasing temperature. There is apparently a considerably greater restriction to rotation in the borneol and isoborneol than in the bornyl halides.

Thiocamphor behaves very much like camphor at the higher temperatures but has a much less sharp transition which occurs at a slightly higher temperature. The curve for thioborneol is almost an exact duplicate of that for bornyl chloride. However, it does not show a transition but the dielectric constant decreases with anomalous dispersion at about -120° . Three of these compounds have their transitions at lower temperature than that of camphor, but in the borneols it is much higher.

This is not the complete list of compounds related to camphor which show molecular rotation but it covers the most important ones. Most of the materials discussed thus far have transitions above which the dielectric constant is high at all frequencies up to 100 kilocycles, the upper limit of the measurements. Below the transition the dielectric constant is practically equal to the square of the refractive index, showing that there is no appreciable contribution from dipole rotations.

Not all of the camphor compounds show this property of molecular rotation in the solid state and so there is some other requirement than simple membership in a certain chemical class. Briefly this additional, and probably most important, requirement is symmetry. Unless the molecule is sufficiently symmetrical, molecular rotation will not be possible in the solid state. This is nicely illustrated by certain camphor derivatives. The change from camphor to chlorocamphor results in a shift of the center of gravity of the molecule away from the center of figure. In chlorocamphor, however, the molecules can still rotate in the solid state but there is one important difference. In camphor the molecules can rotate at relatively low thermal energies, that is, at all temperatures above about -30° , but in chlorocamphor this rotation can take place only above $+75^\circ$. The change from chlorocamphor to bromocamphor results in a still larger shift of center of gravity from the center of figure and in

bromocamphor rotation is not possible below the melting point. An equally striking example of such a change is the difference between camphor and fenchone, which is isomeric with it. Fenchone is a liquid freezing at about $+5^\circ$ as against $+176^\circ$ for camphor; this itself is one evidence of lack of symmetry in fenchone. Although its dielectric constant in the liquid is comparable with that of camphor at the same temperature, on freezing the dielectric constant of fenchone decreases to less than three.

From another and different class of compounds we also have a number of examples of molecules which rotate in the crystalline solid

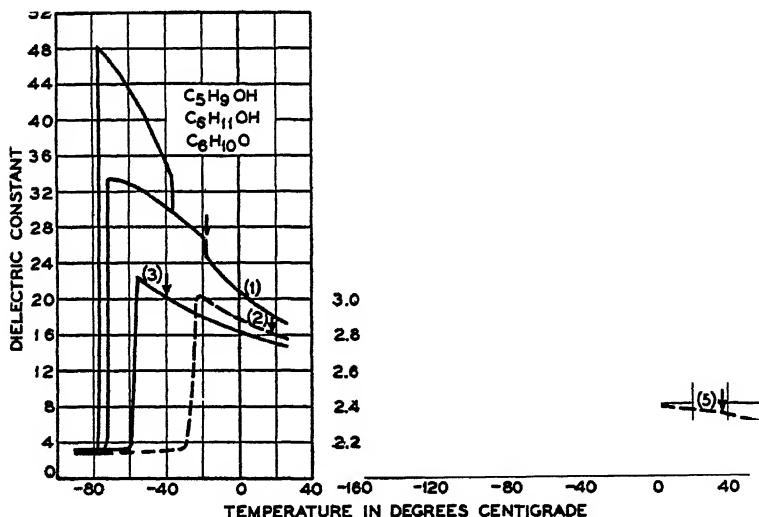


FIGURE 3. Dielectric behavior of cyclopentanol (1), cyclohexanol (2), cyclohexanone (3), 1, 3 cyclohexadiene (4) and camphene (5).

state. FIGURE 3 shows some polar derivatives of cycloparaffins, cyclopentanol, cyclohexanol and cyclohexanone. The data for cyclopentanol are of interest because of the appearance of a second solid transition similar to that of bornyl bromide. Phase II, however, appears to be metastable and is obtained only on cooling. With heating, the molecules start rotating at about -72° but the dielectric constant rises only to a curve which is a continuation of that for phase I.

It is of interest to note that the values of the dielectric constant for solid cyclopentanol are strikingly large as contrasted to those for borneol and isoborneol. They are comparable with those of methyl

and ethyl alcohol and considerably higher than the values for normal amyl or hexyl alcohols. The hindrance to rotation must be small in both the liquid and solid state.

In the same figure, but using a different scale, are the curves for 1, 3 cyclohexadiene, which is polar only because of the unsymmetrical distribution of the double bonds, and of camphene, which has one double bond and also an unsymmetrical distribution of methyl groups. Both show small transitions, that of the cyclohexadiene having a hysteresis.

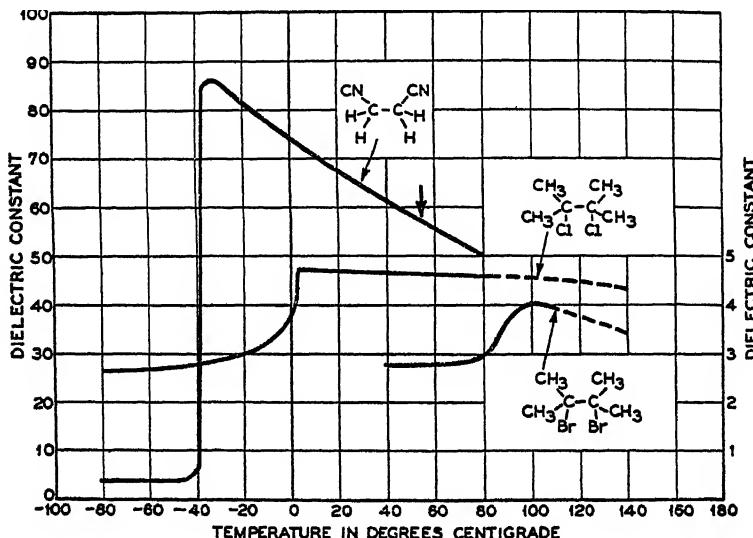


FIGURE 4. Dielectric behavior of ethylene cyanide (1), tetramethyl ethylene chloride (2), and tetramethyl ethylene bromide (3). Use right-hand scale for curves (2) and (3).

Still another group of compounds in which molecular rotation appears to be possible in the crystalline state is the substituted ethanes. These are of interest because of the possibility of rotation about the C-C bond and the possible effect of this rotation upon the dipole moment and ability of the molecule to rotate in the solid state. The most striking of these compounds is ethylene cyanide. It has a high dielectric constant in the liquid state as is shown in FIGURE 4 and thus there is evidence that the molecule is not in the trans form. The value of the electric moment as measured by Williams² also indicates that it is not entirely in the cis form and so it is probably changing its form due to rotation about the C-C bond. The observed moment

² Williams, J. W. Zeit. physikal. Chem. A138: 75. 1928.

corresponds almost to that calculated on the assumption that the molecule is continually changing from cis to trans. The dielectric constant increases almost perfectly continuously through the melting point and has a transition at -40° . The related compound ethylene dichloride has been extensively studied and it does not rotate in the solid state although its moment value and change of moment with temperature indicate considerable motion about the C-C bond at higher temperatures. Similarly with ethylene bromide. We have found, however, that tetramethyl ethylene dichloride and tetramethyl ethylene dibromide do rotate and the curves are shown in FIGURE 4. Whether or not the ability of the molecule to rotate in the solid state depends upon the ease with which it can undergo internal rotation is a question for the theorists. Perhaps in the case of the tetramethyl dihalides and ethylene cyanide the whole molecule is not rotating but we are merely observing the restricted rotation of the parts of the molecule. The data for the tetramethyl dihalides are not complete enough to show whether or not there is a continuity of dielectric constant at the melting point. The ethylene cyanide does, however, show a practically continuous curve of dielectric constant through the melting point.

In a molecule having such freedom of motion about the C-C bond it is very tempting to set up a picture of rotation of parts of the molecule which differs from rotation of the whole molecule only in the matter of time. For example, one part of the molecule may move over a potential barrier to a new position and then a little later the other part move to resume some average equilibrium configuration of the two parts. The result of both motions is the same as though the molecule as a whole had moved. If, then, these motions of the parts of the molecule take place very rapidly as compared to the frequency of the measuring field, the effect is the same as though the molecule as a whole were rotating. When we speak of the molecule rotating it is probably some such process as this that we are thinking of and not a free unrestricted rotation of a rigid molecule. If this picture be valid it is easy to see how molecules which have the freedom of motion of camphor, cyclohexanol or ethylene cyanide should be able to rotate in the solid state and have high dielectric constants while more rigid ones cannot. Also it is easy to understand why putting a long side chain on such a molecule should make it more difficult or even impossible for it to rotate, which has been our experience.

The next class of substances in which molecular rotation has been observed is the aromatics or derivatives of benzene. Nitrobenzene and chlorobenzene have been extensively studied in both the liquid and solid state and it is quite certain that both have low dielectric constants in the solid, showing that the rotation which takes place in the liquid is impossible in the solid. Most of the benzene derivatives which have been measured heretofore have behaved in this way and if it is spherical symmetry or freedom of motion about C-C bonds which is required, then that is quite the expected behavior. However, in

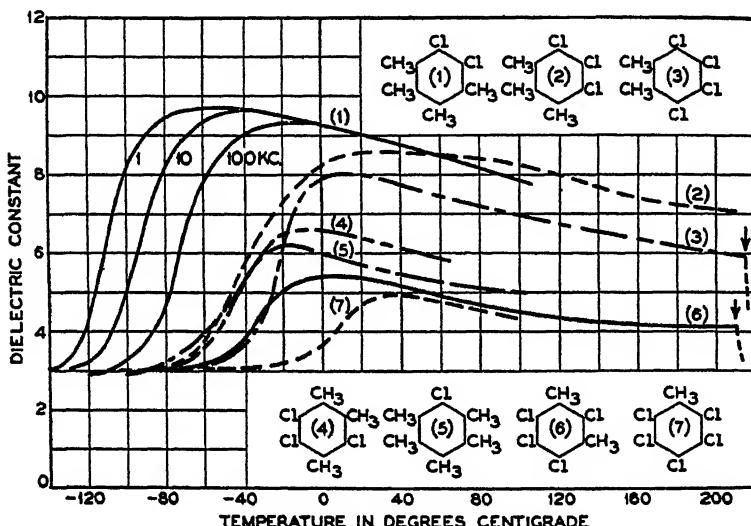


FIGURE 5 Dielectric behavior of hexasubstituted chloromethyl benzenes at 100 kc; structure indicated on curve.

considering the rotation of molecules from the standpoint of symmetry it appeared that for benzene derivatives we should strive for circular symmetry. If we achieve that symmetry by substituting six of the same groups in benzene we have no polarity and so changes in rotation do not directly affect the dielectric constant. By reason of the difference in moment between a CH_3 group and Cl, however, and their approximately equal size, we can by a suitable combination of substituents attain both polarity and reasonably good geometrical symmetry.

FIGURE 5 shows the results of measurements on a number of hexa-substituted chloromethyl benzenes. All but one of the possible polar combinations of CH_3 and Cl completely substituted benzenes are

shown on the curve. All of these compounds have high melting points, about 200°, as is to be expected for symmetrical compounds, and in the few cases where measurements have been made in the liquid as well, the dielectric constant has been shown to increase on solidifying.

All of these compounds show molecular rotation. However, this rotation drops out with anomalous dispersion instead of at a transition as in the camphor, cycloparaffin and ethane derivatives. As the temperature decreases the frequency with which the molecule can turn through some angle and pass over a barrier into the next valley becomes less and finally when it becomes less than that of the measuring field, as it does at lower temperatures, the dielectric constant begins to decrease and approaches a value of about three, which is approximately equal to the square of the refractive index.

The curves for three frequencies are shown for one material, the dichloroprehnitene, while for the other isomers the curves are all for 100 kilocycles. In the lower group of compounds there is one effective Cl vector moment, the others cancelling each other. For these compounds the dielectric constants at room temperature range between 5 and 6, which compares with the value 5.5 for liquid monochlorobenzene at the same temperature. This provides food for thought about the extent of hindrance in the crystalline solid as compared to that of the liquid. The two compounds having two adjacent effective Cl vectors directed at 60° angles have dielectric constants ranging from 8 to 9; that for liquid orthodichlorobenzene is 10. The 1, 2, 3 trichloro, 4, 5, 6 trimethyl benzene, which should have a slightly higher moment than dichloroprehnitene, for example, shows a slightly higher dielectric constant at high temperatures but it shows a greater interaction reducing the dielectric constant at lower temperatures.

Thus it seems that these hexasubstituted benzenes are sufficiently symmetrical for the molecules to rotate in the solid state. It seems almost necessary that such rotation be about an axis perpendicular to the plane of the ring; the symmetry is decreased about any other axis by substitution. The effect of the substitution of other groups in benzene has been only very sketchily investigated. It appears clear, however, that a longer group such as ethyl or a nitro group which would extend out of the plane reduces the likelihood of rotation.

During this work some pentasubstituted chloromethyl benzenes were prepared and they were also found to rotate; thus complete symmetry is not essential. However, the tetra and lower substituted

compounds which have been prepared do not rotate. FIGURE 6 shows data for three of these pentasubstituted chloromethyl benzenes. They differ from the hexasubstituted compounds in that they show the dielectric constant decreasing at a transition more like the camphor or cycloparaffin compounds instead of with anomalous dispersion. In these less symmetrical compounds it seems likely that this is because of the possibility of a collapse to an interlocking and more densely packed form. This is indicated also by measurements of

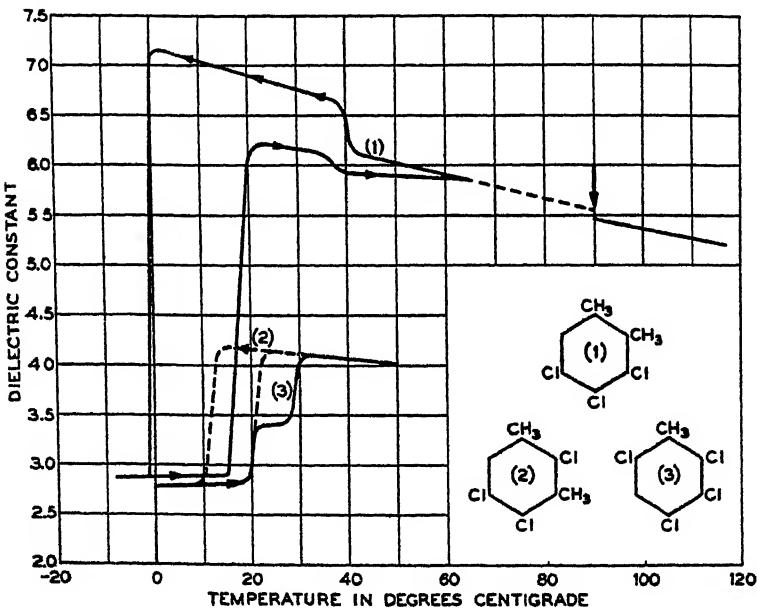


FIGURE 6. Dielectric behavior of pentasubstituted chloromethyl benzenes. Structure indicated on curve. Arrows on curve indicate direction of temperature change where hysteresis was observed.

molecular volume, which show that below the transition the pentasubstituted compounds have a lower value than that of the hexasubstituted compounds. Above the transition point, *i. e.* when the molecules are rotating, the molecular volume of the pentasubstituted compounds increases to approximately the same value as that of the hexasubstituted compounds.

The picture has been simplified somewhat in order to make it possible to cover such a large number of different compounds. Many of them show some peculiarity such as a hysteresis or effect of frequency but it does not seem possible that full consideration of these

would change the essential picture, which is that large organic molecules do rotate in the crystalline solid state. Some, like camphor, because they approach very closely to a sphere in their structure and therefore have low fields opposing rotation. Others, such as cyclohexane derivatives and ethylene cyanide, perhaps because their freedom of motion makes it easy for them to wriggle or flop their way around or possibly because this motion endows them with a dynamic symmetry. And finally, the benzene compounds, which can rotate because of symmetry about a single axis.

The physical properties of these materials are strikingly in contrast (except for the hexasubstituted benzenes) to those for so-called normal organic crystals in which molecules are not rotating. They are almost all soft and waxy and have high vapor pressures, which is to be expected if they have very low internal fields. It has been known for a long time that camphor had an anomalously high freezing point depression constant. Pirsch has shown that also to be the case for many of the other compounds which dielectric evidence shows to be capable of rotating. Unlike the normal organic materials they pass from the liquid to the lower energy solid state in two steps, losing only a part of their energy at the fusion point and the rest at the transition point. The high freezing point constant is a consequence of this low heat of fusion and like the high dielectric constant of the solid is traceable to the symmetry of the molecules.

THE DIELECTRIC PROPERTIES OF PROTEIN SOLUTIONS

BY J. L. ONCLEY, J. D. FERRY, AND J. SHACK

*From the Department of Physical Chemistry, Harvard Medical School, Boston,
Massachusetts*

The behavior of a dipole molecule in the frequency region where $2\pi\nu\tau$ is of the order of unity, τ being the so-called "relaxation time" of the molecule and ν the frequency, was first discussed by Debye.¹ Several systems showing this phenomenon of anomalous dispersion have been investigated, and their behavior is at least approximately as predicted by the equations of Debye. In recent years advances in the technique of dielectric measurements upon conducting solutions and in the interpretation of results in polar solvents have made possible the systematic investigation of a number of protein solutions, and interpretation of the data in terms of the Debye theory. The results obtained are of considerable interest in both the theory of dielectric dispersion and the study of the fundamental structure of proteins.

THEORY

It is hardly necessary to review the fundamental aspects of the dielectric dispersion of large molecules in non-polar solvents. When dealing with polar solvents, as we must in the case of all protein solutions, we introduce certain empirical relations of Wyman,² supported by the calculations of Onsager.³ The typical behavior of a simple binary mixture involving two relaxation times of widely differing magnitude is illustrated in FIGURE 1. The three quantities plotted are defined as follows:

1. Dielectric constant (real part); $\epsilon' = C/C^*$
 2. Specific conductivity; $\kappa = 0.0885 G/C^*$
 3. Dielectric absorption; $\Delta\epsilon'' = (G - G_0)/(2\pi\nu C^*) = 1.80(\kappa - \kappa_0)/\nu$
- Here C and G represent the capacitance in $\mu\mu\text{fds}$ and conductance (parallel) in μmhos of the cell filled with solution (the effect of lead capacitance being eliminated), C^* the change in cell capacitance caused by a unit change in the dielectric constant, ν the frequency in megacycles, and G_0 (and κ_0) the conductance (and specific conduc-

¹ Debye, P. "Polar Molecules." Chem. Cat. Co., New York. 1929.

² Wyman, J. Jour. Am. Chem. Soc. 58: 1482. 1936.

³ Onsager, L. Jour. Am. Chem. Soc. 58: 1486. 1936.

tance) of the solution at very low frequencies in μ mhos. The figure is divided into five regions. In region *A* the frequency of the alternating voltage is sufficiently low to allow orientation of dipoles of both types. In region *B*, however, the orientation of the dipoles of larger relaxation time lags, and is no longer independent of the frequency, giving a region of decreasing dielectric constant and increasing specific conductance. The dielectric absorption goes through a maximum in this region and the frequency of this maximum (also of mean

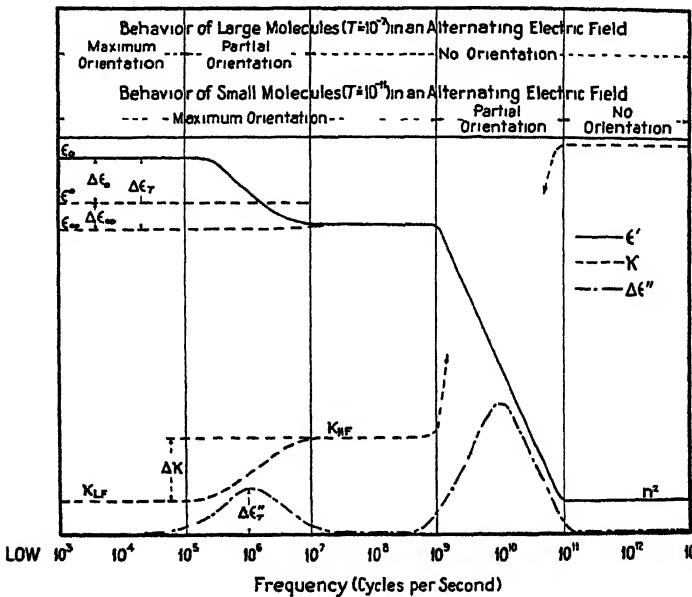


FIGURE 1. Schematic diagram of anomalous dispersion of the dielectric constant (ϵ'), the specific conductivity (κ), and the dielectric absorption ($\Delta\epsilon''$), for two widely separated critical frequencies. (Shack, Ph.D. Dissertation, Harvard University 15. 1939.)

dielectric constant and conductance) characterizes the region and is called the critical frequency. In region *C* the frequency is so great that there is no orientation of the larger molecules, and we have a region of constant and intermediate dielectric constant and specific conductance and practically no dielectric absorption. Regions *D* and *E* indicate similar behavior of the smaller molecules.

The equations representing the behavior in region *B* are:

$$\begin{aligned} \epsilon' &= \epsilon_0 - (\epsilon_0 - \epsilon_\infty) v^2/(v^2 + v_c^2) = \epsilon_\infty + (\epsilon_0 - \epsilon_\infty) v_c^2/(v^2 + v_c^2) \\ \Delta\epsilon'' &= (\epsilon_0 - \epsilon_\infty) vv_c/(v^2 + v_c^2) \quad (1) \\ \kappa &= \kappa_0 + (\kappa_\infty - \kappa_0) v^2/(v^2 + v_c^2) = \kappa_\infty - (\kappa_\infty - \kappa_0) v_c^2/(v^2 + v_c^2) \end{aligned}$$

If more than one relaxation time be present, these equations can be written in the more general form:

$$\begin{aligned}\epsilon' &= \epsilon_0 - \Delta\epsilon_{t_1}v^2/(v^2 + v_{c_1}^2) - \Delta\epsilon_{t_2}v^2/(v^2 + v_{c_2}^2) - \dots \\&= \epsilon_\infty + \Delta\epsilon_{t_1}v_{c_1}^2/(v^2 + v_{c_1}^2) + \Delta\epsilon_{t_2}v_{c_2}^2/(v^2 + v_{c_2}^2) + \dots \\ \Delta\epsilon'' &= \Delta\epsilon_{t_1}vv_{c_1}/(v^2 + v_{c_1}^2) + \Delta\epsilon_{t_2}vv_{c_2}/(v^2 + v_{c_2}^2) + \dots \\ \kappa &= \kappa_0 + \Delta\kappa_{t_1}v^2/(v^2 + v_{c_1}^2) + \Delta\kappa_{t_2}v^2/(v^2 + v_{c_2}^2) + \dots \\&= \kappa_\infty - \Delta\kappa_{t_1}v_{c_1}^2/(v^2 + v_{c_1}^2) - \Delta\kappa_{t_2}v_{c_2}^2/(v^2 + v_{c_2}^2) - \dots\end{aligned}\quad (2)$$

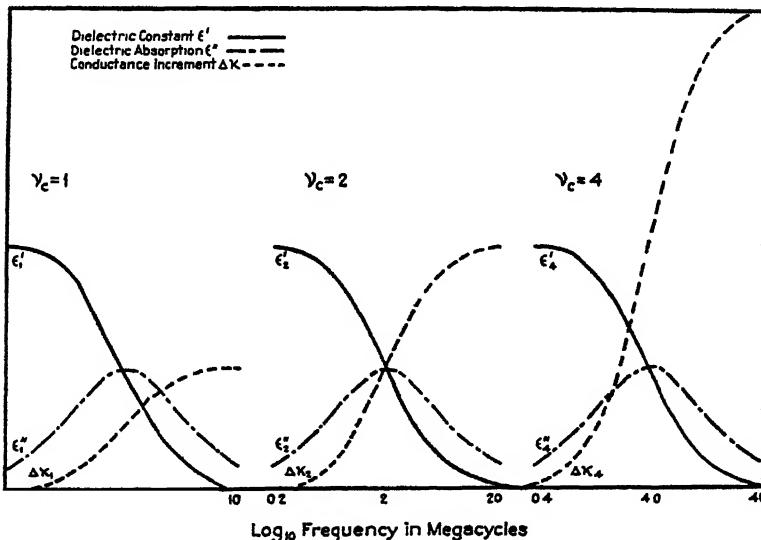


FIGURE 2. Comparative magnitudes of dielectric and conductance increments for dispersions at different critical frequencies. (Shack, Ph.D. Dissertation, Harvard University 12. 1939.)

Here v_{ci} , $\Delta\epsilon_{ti}$, and $\Delta\kappa_{ti}$ are the critical frequency, the dielectric increment, and the conductance increment of the i th dispersion region. There is always a relation between $\Delta\epsilon_{ti}$ and $\Delta\kappa_{ti}$ of the form

$$\Delta\kappa_{ti} = \Delta\epsilon_{ti} v_{ci} / 1.80. \quad (3)$$

This relation gives us the important fact that dispersion regions with large values of v_{ci} (high critical frequencies) will give larger conductance increments for equal dielectric increments. FIGURE 2 shows this fact in graphic form.

Relaxation times τ_i are calculated from the critical frequencies by means of the relation

$$\tau_i = 1/(2\pi v_{ci}). \quad (4)$$

These relaxation times are intimately connected with the size and shape of the molecule. Perrin has derived equations for relaxation times for ellipsoids of revolution of both elongated and flattened nature, to be referred to later. The dielectric increment $\Delta\epsilon_t$ is directly related to the polarizability of the molecule, and this in turn to the electric moment.

METHODS

The measurements of dielectric constants have been made by both resonance and bridge methods,^{4, 5} the bridge being used for frequencies below 2.5 megacycles and the resonance method for higher frequencies. The measurements of conductance have been carried out with a calorimetric apparatus described elsewhere.⁶ The solution to be studied is contained in a dilatometer; the heating produced by the applied field causes an expansion of the solution, which is measured by observing the rate of rise of the level in the capillary. The specific conductance is proportional to this rate of rise and the reciprocal of the square of the applied voltage. The proportionality constant depends upon the cell constant, the radius of the capillary, the density, the coefficient of cubic expansion, and the specific heat, these three latter quantities varying slightly for different solutions. It can be determined by measurement of the rate of rise at a very low frequency where the specific conductance can be measured with an ordinary conductance bridge.

RESULTS

A survey of the dielectric constant literature for protein solutions reveals few data of significance before those of Wyman⁷ on the plant protein zein. Up to this time most workers had made measurements at such high frequencies that the results obtained fell in region C (FIGURE 1) where the contribution of the protein molecules to the dielectric effect was very small; and even these measurements were very uncertain. The development of new techniques of measurement, together with the low conductance of zein solutions (zein has a very small number of dissociable groups and dissolves in about 70% ethyl or propyl alcohol) made it possible to observe for the first time the dispersion region, and to obtain extrapolated values for ϵ_0 which

⁴ Wyman, J. Phys. Rev. 35: 623. 1930.

⁵ Oncley, J. L., Ferry, J. D., & Shack, J. Cold Spring Harbor Symposia Quant. Biol. 8: 21. 1938.

⁶ Shack, J. Ph.D. Dissertation, "Dielectric Absorption of Protein Solutions," Harvard University. 1939.

⁷ Wyman, J. Jour. Biol. Chem. 90: 443. 1931.

could be considered as a measure of the polarizability and dipole moment of the protein molecule. The results obtained thus indicated that to a first approximation, at least, one could use the Debye dispersion theory to explain the dielectric measurements, and that the relaxation times and dipole moments were of a reasonable order of magnitude. Since these original measurements of Wyman in 1931, results on a considerable number of proteins have been obtained in several laboratories, and the measurements have definitely progressed beyond the first approximation. These results have been contributed largely at first by Errera⁸ and Marinesco,⁹ and currently by Andrews, Arrhenius, Elliott, Ferry, Oncley, Shack, and Williams. Measurements have also been made at low frequencies by Shutt.¹⁰

From an experimental viewpoint the proteins can be divided into several classes: first, those soluble in alcohol-water mixtures and of relatively low conductance, called the prolamines (zein, gliadin and secalin); second, those sufficiently soluble in water without the addition of salt, which have a considerably larger conductance (egg albumin, serum albumin, hemoglobin, serum pseudoglobulin, and the slightly soluble lactoglobulin); third, those which can be dissolved in certain solvent mixtures of water, urea, glycine, ethylene glycol, propylene glycol, glycerol, etc. This third class is being investigated in our laboratory at the present time, and is represented by insulin in this review. Solutions of those proteins requiring the addition of salt before becoming soluble are of too high a conductance to be measured at the present stage of development.

Hemoglobin

Perhaps the simplest dielectric behavior we have observed for any protein is that of hemoglobin. Measurements of Oncley on crystalline CO-hemoglobin (horse)^{11, 12} showed that both the low frequency dielectric increment $\Delta\epsilon_0$ and the high frequency increment $\Delta\epsilon_\infty$ are linear functions of the concentration, and that the dispersion curves are characterized by a single critical frequency (and hence relaxation time). These results are reproduced in FIGURES 3 and 4 (curve 3). Measurements by Shack⁶ on the conductance increment are in agreement with these observations and are reproduced in FIGURE 5.

⁸ Errera, J. *Jour. Chim. Phys.* 29: 577. 1932.

⁹ Marinesco, N. *Kolloid Zeit.* 58: 285. 1932.

¹⁰ Shutt, W. J. *Trans. Faraday Soc.* 30: 893. 1934.

¹¹ Oncley, J. L. *Jour. Am. Chem. Soc.* 60: 1115. 1938.

¹² The hemoglobin of the pig has been studied by Arrhenius, S. *Physik Zeit.* 39: 559. 1938.

The curves in FIGURES 4 and 5 represent the values calculated from equations (1) and no significant deviations are observed. The numerical values for critical frequencies and increments obtained by these two methods are shown in a later table.

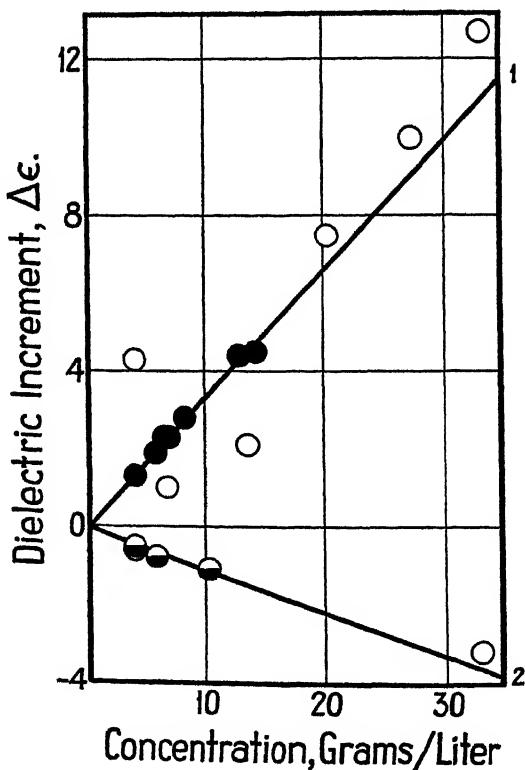


FIGURE 3. Low (1) and high (2) frequency dielectric increments of carboxyhemoglobin: o, Errera⁸; ●, Wyman¹¹; •, Oncley¹¹. (Oncley, Jour. Am. Chem. Soc. 60: 1121. 1938.)

Serum Albumin

The behavior of crystalline serum albumin fractions (horse) of various solubilities is only slightly more complicated.¹³ It was shown that preparations whose solubility in ammonium sulfate solutions varied by a factor of about five had low frequency dielectric increments varying from 0.12 dielectric constant units per gram per liter for the least soluble to 0.39 for the most soluble. This is shown by curves 2 to 5 in FIGURE 6. When reduced to the same scale

¹¹ Ferry, J. D., & Oncley, J. L. Jour. Am. Chem. Soc. 60: 1123. 1938.

these preparations were all found to give identical dispersion curves, as indicated by curve 2 in **FIGURE 4**. This result was of considerable importance from the standpoint of protein chemistry, since it gave a new method for differentiation of the fractions with varying solubility but almost identical size and charge.

Lactoglobulin

Solutions of lactoglobulin in water (about 0.1% concentration) or in glycine solutions of various concentrations show the same

Dispersion Curves of Blood Proteins (Horse)

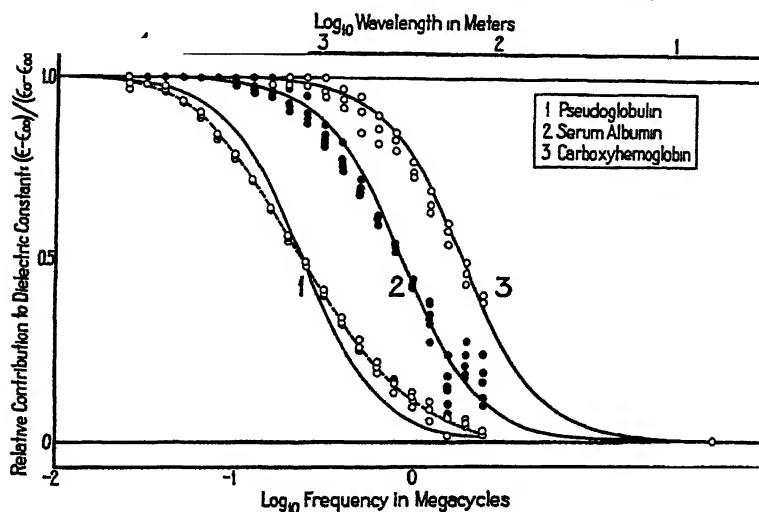


FIGURE 4. Dispersion of the dielectric constant of blood proteins of the horse: 1, serum pseudoglobulin¹⁴; 2, serum albumin¹⁵; 3, carboxyhemoglobin¹¹. (Cohn, Chem. Rev. 24: 217. 1939.)

simple behavior as observed for hemoglobin and serum albumin.¹⁴ The results obtained by the calorimetric method⁶ are similar, except for a very small component of high critical frequency (12 megacycles).

Insulin

Insulin, which is almost insoluble in water at the isoelectric point, has been studied in mixtures of water and propylene glycol.¹⁵ The

¹⁴ Ferry, J. D., Cohn, E. J., Oncley, J. L., & Blanchard, M. H. Jour. Biol. Chem. 128: Proc. 28: 1939.

¹⁵ Cohn, E. J., Ferry, J. D., Livingood, J. J., & Blanchard, M. H. Science 90: 183. 1939.

Dielectric Absorption of Carboxyhemoglobin Solutions.

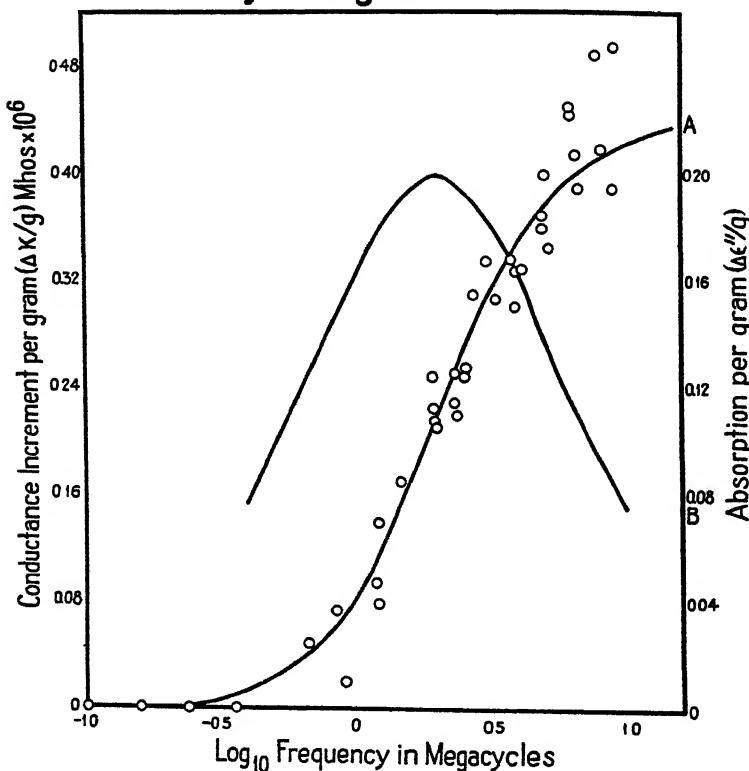


FIGURE 5. Dielectric absorption of carboxyhemoglobin solutions: o, experimental measurements of $(\kappa - \kappa_0)/g$; curve A, $(\kappa - \kappa_0)/g$, and curve B, $\Delta\kappa''/g$, calculated from equation (1) using the constants recorded in TABLE I for the calorimetric method (Shack, Ph.D. Dissertation, Harvard University 50. 1939.)

dielectric increment varies somewhat with composition of solvent, increasing with increasing water content. The dispersion curves, when analyzed for single critical frequencies, yield relaxation times which are directly proportional to the viscosities of the respective solvents (within 10%). They are shown in FIGURE 7.

Zein¹⁶

The results on certain other protein solutions cannot be explained on the basis of a single critical frequency (and hence single relaxa-

¹⁶ The prolamines gliadin and secalin have also been studied. See Arrhenius, S. Jour. Chem. Phys. 5: 63. 1937. Williams, J. W., & Watson, C. C. Cold Spring Harbor Symposia Quant. Biol. 6: 203. 1938. Andrews, Dissertation Univ. of Wisconsin. 1938.

tion time). The measurements of Elliott and Williams¹⁷ on zein are typical of this group of proteins. FIGURE 8, taken from their paper, shows how two Debye curves may be used in the interpretation of these data.

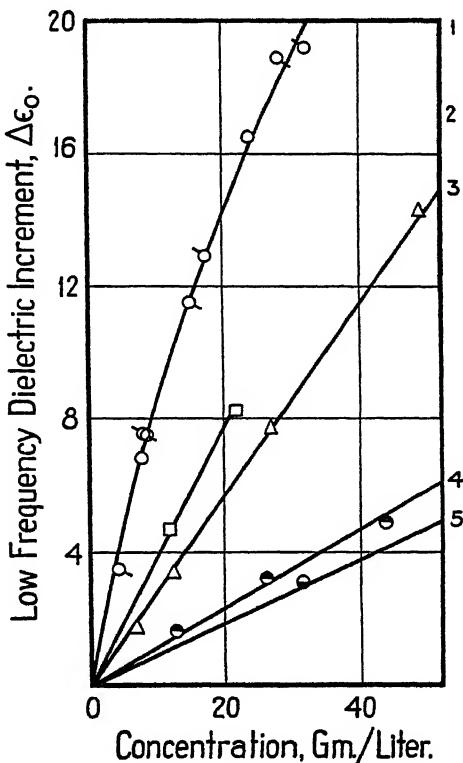


FIGURE 6 Low frequency dielectric increments of horse serum proteins¹²: 1, pseudoglobulin; 2-5, albumins of successively decreasing solubilities in concentrated ammonium sulfate solutions. (Ferry & Oncley, Jour. Am. Chem. Soc. 60: 1129. 1938.)

Egg Albumin

Egg albumin solutions are observed to exhibit effects similar to those described from zein solutions.^{6, 10, 18} Here the experimental determination of the dispersion data is much more difficult, however, since the dielectric constant increment is small, the solutions

¹⁷ Elliott, M. A., & Williams, J. W. Jour. Am. Chem. Soc. 61: 718. 1939. See also Williams, J. W., & Watson, C. C. Cold Spring Harbor Symposia Quant. Biol. 6: 208. 1938.

¹⁸ Oncley, J. L. Unpublished data.

Dispersion of the Dielectric Constant of Insulin Solutions

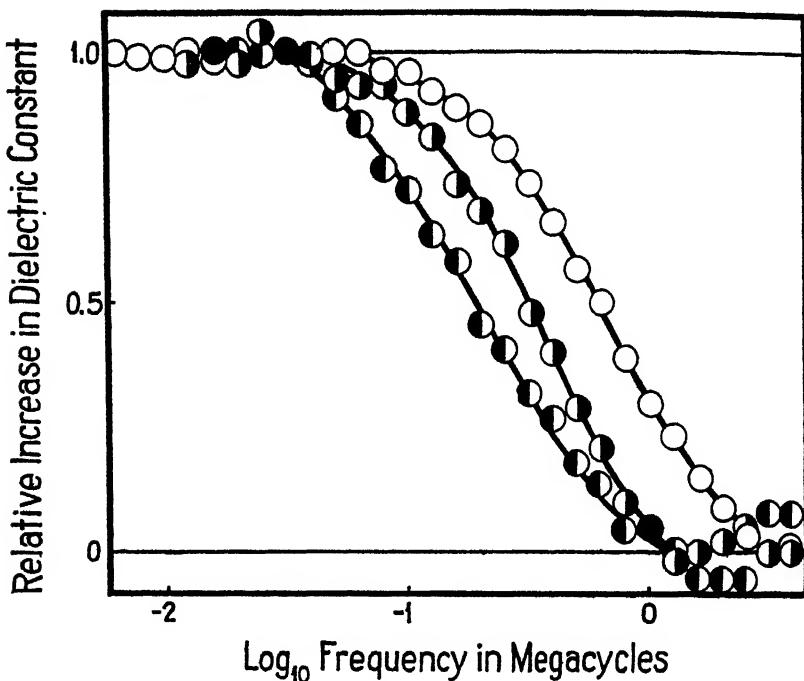


FIGURE 7. Dispersion of the dielectric constant of insulin ¹¹ in (O), propylene glycol (viscosity relative to water, $\eta_r = 47.5$); (●), propylene glycol with 10% water ($\eta_r = 28$); (◐), propylene glycol with 20% water ($\eta_r = 17$). (Cohn, Ferry, Livingood, & Blanchard, Science 90: 184. 1939.)

are much more conducting, and the two dispersion curves are much closer together and hence more difficult to resolve. The results given here are for a fairly concentrated solution of about 9% egg albumin. More dilute solutions are being studied. FIGURE 9 shows the dispersion curve of such a solution, resolved into two Debye curves. The broken curve in this figure represents values calculated by means of equation (2) from conductances measured independently by the calorimetric method.

Serum Pseudoglobulin

Similar results were also obtained by Ferry and Oncley^{12, 13} in the case of serum pseudoglobulin solutions. Curve 1 in FIGURE 4 il-

¹² In this paper, Ferry, J. D., & Oncley, J. L. stated that the observed dispersion could be represented by a sum of two terms. Several possible interpretations of these two

lustrates this result, the heavy curve representing the result calculated from equation (1) for a single critical frequency, and the broken curve representing the sum of two terms of equation (2). The low frequency dielectric increments $\Delta\epsilon_0$ are plotted as a function of the concentration in FIGURE 6, curve 1, and show the same type of behavior as described for hemoglobin and serum albumin, except that the curve deviates from a linear relationship at the higher concentrations—an effect that may be related to the very high dipole moment of the pseudoglobulin molecule.²⁰

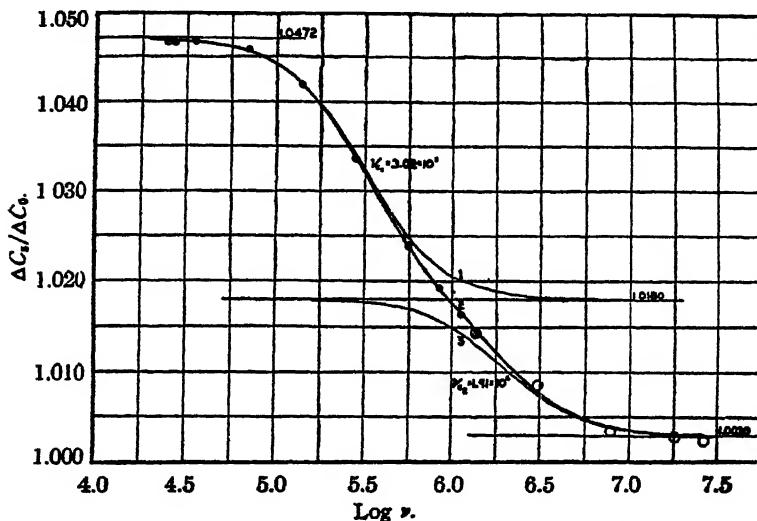


FIGURE 8. Dispersion of the dielectric constant of zein¹⁷, showing resolution into two Debye curves. (Elliott & Williams, Jour. Am. Chem. Soc. 61: 721. 1939.)

terms were discussed; thus, on p. 1130, they ". . . might be ascribed either to rotation of the same asymmetrical molecule about different axes, or to the presence of two or more symmetrical molecular species." In the absence of information concerning the molecular weight of the pseudoglobulin preparations employed, no attempt was made to eliminate either of these possibilities. It was, however, pointed out that the calculated and observed relaxation times did not agree if the scanty data available were used with the interpretation involving two molecular species. This statement was presumably misunderstood by Elliott, M. A., & Williams, J. W. (Jour. Am. Chem. Soc. 61: 724. 1939), who write ". . . the attempt was made to associate the two times of relaxation with two different molecular units of weight of 175,000 and 400,000 which were assumed to be present. In other words, the question of molecular asymmetry is not involved here." Further investigations on pseudoglobulin preparations of known sedimentation constant, diffusion constant, and electrophoretic mobility are in progress; from these it should be possible to distinguish between the above hypothetical interpretations of the dielectric dispersion curves. The investigations referred to here have been discussed by Oncley, Jour. Phys. Chem., in press. The first interpretation of Ferry and Oncley, namely that the behavior ". . . might be ascribed . . . to rotation of the same asymmetrical molecule about different axes . . ." has been borne out by these new data.

²⁰ Cf. Fuoss, R. M. Jour. Am. Chem. Soc. 56: 1031. 1934; 58: 982. 1936.

of both parts of the complex dielectric constant throughout the dispersion region for protein solutions. The mean values in this table are weighted, more weight being given to the low frequency bridge data and to the high frequency calorimetric data. Values enclosed in parentheses are calculated from equation (3). FIGURE 9 also compares the data obtained by these two methods in the case of egg albumin.

Relaxation Time

The values of the relaxation times for practically all proteins which have been investigated in a comprehensive fashion are recorded in columns two and three of TABLE 2. The last four proteins recorded there have yielded two relaxation times, the others only one. If we assume that the protein molecule may be represented by an ellipsoid of revolution, the treatment of Perrin²¹ may be used to reduce the two observed relaxation times τ_a and τ_b to values of the relaxation time of a sphere of equal volume, τ_0 , and the axial ratio a/b . Here a is taken as the half-length of the ellipsoid along the geometric axis of rotation, and b as the equatorial radius. These values are recorded in columns four and six of TABLE 2. The values τ_0' column five, are calculated from the molecular volumes of these proteins as obtained from a combination of sedimentation velocity and diffusion measurements,²² using the equation $\tau_0' = 3\eta V/RT$, η being the viscosity of water at 25° C, V the molecular volume, T the absolute temperature, and R the gas constant (8.31×10^7 ergs per °C per mole). When only one relaxation time has been observed, calculations of this type cannot be made, and the asymmetry ratio a/b is obtained from the observed and the calculated relaxation times. The absence of a second relaxation time might be caused by any of at least three things; (1) the relaxation time may be so large as to fall in a frequency region not studied, (2) the magnitude of the dispersion may be so small as to make its measurement very difficult, and (3) the two relaxation times may be very nearly the same. The second of these suggestions is the most likely, since at present a dispersion would have to amount to at least 10 or 15% of the total dispersion if it were to be observed with much certainty. In cases where only one relaxation was observed, therefore, it is possible that this rotation may have involved either the long or the short geometric axis,

²¹ Perrin, F. Jour. Phys. Radium. 5: 497. 1934.

²² These molecular volumes are in some cases incompatible with the molecular weights given in column ten, TABLE 2, which are weighted means of the molecular weights as obtained by various methods. (Cohn, E. J. Chem. Rev. 24: 203. 1939.)

TABLE 2
RELAXATION TIME, ASYMMETRY, AND DIPOLE MOMENT OF PROTEIN MOLECULES WHEN TREATED AS ELONGATED ELLIPSOIDS OF REVOLUTION

Protein	Relaxation Time ($\times 10^6$) reduced to water at 25°						Asymmetry			Total Dielectric Increment $\Delta \epsilon_{1/g}$	Assumed Molecular Weight M	Dipole Moment (Debye Units) μ	Dipole Moment Angle Θ
	Long Axis r_a	Short Axis r_b	Equiv. Sphere	Calc.	Dielectric a/b	Diffusion a/b	Viscosity a/b						
	τ_a	τ_b	τ_0	τ_0'									
Carboxyhemoglobin (horse) ¹¹	8.4 ^f	—	—	5.6	2 ^f	5	6	0.40	66,700	480	—	—	—
Carboxyhemoglobin (pig) ¹²	17. ^f	—	—	(5.6)	4 ^f	—	—	.3	(66,700)	400	—	—	—
Serum Albumin (horse) ¹³	19. ^f	—	—	5.7	4 ^f	5	7	.18	73,000	{280}	{510}	—	—
Lactoglobulin ^{6, 14, e}	7.0 ^f	—	—	3.4	3 ^f	5	7	1.5	40,000	700	—	—	—
Insulin ^{6, f}	1.7 ^a	—	—	2.8	— ^a	3	—	0.38	35,000	330	—	—	—
Gliadin ^{6, e}	—	3.7 ^h	—	2.6	— ^a	11	—	.15	42,000	230	—	—	—
Zein ^{7, d}	24.1	4.2	3.3	3.3	7	14 ^f	—	.45	39,000	380	35°	—	—
Seasolin ^{6, e}	29.0	2.4	1.9	—	11	—	—	1.0	24,000	440	40°	—	—
Egg Albumin ^{6, 15}	18.5	4.7	3.7	3.5	5	4	6	0.17	34,000	210	40°	—	—
Serum Pseudoglobulin (horse) ¹¹	130.	33.	26.	13.5	5	8	10	1.2	162,000	1200	45°	—	—

^a Dissolved in M/2 glycine. Lactoglobulin has the same dielectric increment and relaxation time when dissolved in water or M/4 glycine.

^b Dissolved in 80% propylene glycol. Insulin has the same relaxation time, but lower dipole moment when dissolved in propylene glycol containing less than 20% water.

^c Dissolved in 60% ethanol. ^d Dissolved in 72% ethanol. ^e Dissolved in 53% ethanol. ^f The asymmetry is calculated on the basis that the observed relaxation time is τ_a . ^g The observed relaxation time is probably τ_b , and no estimate of a/b can be made. If the observed time is taken as τ_a , the value of a/b obtained is 2.

and there is some uncertainty in the calculation of a/b since the one observed relaxation time may be either τ_a or τ_b . There is one restriction, however, in that τ_b/τ_0' cannot be greater than 4/3 according to the equations of Perrin. All of the cases in TABLE 2 where only one relaxation time was observed have values of τ/τ_0' in excess of 4/3, so that any assumption of rotation about the long axis would necessitate the further assumption of sufficient hydration to reduce this ratio to at least 4/3. This seems to be the proper explanation in the case of gliadin, where the ratio is only slightly greater than 4/3 and where the asymmetry is known to be high. In all other cases we have used the observed relaxation time as τ_a in the calculations in TABLE 2. The agreement between the observed and calculated relaxation times for spheres of equivalent volume is very satisfactory in most cases. However the observed relaxation times for serum pseudoglobulin are too large by a factor of almost two; and for insulin too small by a like factor. These discrepancies cannot be adequately explained on any simple basis, and require further investigation.

Asymmetry

The values obtained for the asymmetry, a/b , by the treatment just discussed are recorded in column six of TABLE 2. For comparison we have also recorded asymmetries calculated from diffusion constants, using another equation of Perrin,²³ and from viscosity, using equations of Guth.²⁴ We are indebted to Dr. John W. Mehl for these comparisons. The agreement is quite satisfactory except in a few cases. The different asymmetries of horse and pig hemoglobin need further study. The assumption of a moderate degree of hydration would make slight modifications of the asymmetries. The dielectric method for evaluating the shape of molecules when two relaxation times can be measured represents perhaps the only method as yet available for the calculation of asymmetry which is entirely independent of hydration.

Dipole Moment

The dipole moments of these proteins are recorded in column eleven of TABLE 2. They are calculated from the equation¹¹

$$\mu = 2.9 \sqrt{M \Delta \epsilon / g} \quad (5)$$

²³ Perrin, F. Jour. Phys. Radium. 7: 1. 1936.

²⁴ Guth, E. Kolloid. Zeit. 74: 147. 1936. For more recent calculations of this type we have used the newer equations of Simha, which give somewhat smaller asymmetries. See Mehl, J. W., Oncley, J. L., & Simha, R. Science 92: 132. 1940.

using the values of the total dielectric increment and of the molecular weight recorded in columns nine and ten. These dipole moments are seen to vary over a wide range, and to be of quite a different order of magnitude from known moments of other types of molecules. These large moments are due largely to the high molecular weight of the proteins, however, and really represent a high degree of electrical symmetry in these structures. Thus hemoglobin with at least seventy-five positively and negatively charged groups²⁵ could have a moment of the order of 500 Debye units if only two pairs of charged groups were placed at a maximum distance apart (assuming a spherical shape). In the cases where two terms were observed in the dispersion equations, we may calculate the components of the dipole moment along the long and the short axes; this result can also be expressed in terms of a "dipole angle," Θ , representing the angle between the axis, a , of the ellipsoid and the dipole moment vector. This dipole angle is thus defined by the equation

$$\tan \Theta = \mu_b / \mu_a = \sqrt{\Delta \epsilon t_b / \Delta \epsilon t_a} \quad (6)$$

and is recorded in the last column of TABLE 2. This angle has been measured only in cases where it is in the general region of 35 to 45°, probably because the experimental studies have given resolvable curves only when $\tan \Theta$ is of the order of unity. It would seem rather unlikely that this angle should be much larger than 45° in the case of the more asymmetrical molecules; gliadin may, however, be an exception to this statement.

SUMMARY

Dielectric dispersion data for protein solutions can be obtained with a satisfactory degree of accuracy by two totally different methods, one for capacitance and one for conductance. The interpretation of these measurements in terms of critical frequencies and increments can be made with but little ambiguity. When compared with similar studies on other macromolecules they seem to indicate a characteristic degree of rigidity in protein structures. Absolute values of size, shape, and dipole moment calculated from these measurements will be slightly affected by the application of various equations for the dielectric behavior of protein solutions, but the empirical relations of Wyman as used in the calculations of this

²⁵ Cohn, E. J. Chem. Rev. 19: 241. 1936.

paper cannot give results far in error. The minor discrepancies in the interpretation of certain points must first be checked to eliminate any possibility of experimental errors, and then we may have to attempt to describe the structure of the protein molecule to a higher degree of approximation. It may be that the assumption of elongated ellipsoids of revolution is not of sufficient generality, and that the assumption of fixed charge distribution, and therefore definite permanent moment, must be modified by a more detailed study of the ionization of the various acidic and basic groups of the protein. Furthermore, the development of experimental methods capable of dealing with solutions of higher conductance would be of great value. It is clear, however, that dielectric measurements interpreted on the basis of the simple theory which has been developed are capable of contributing considerably to our knowledge of the size, shape, and electrical symmetry of protein molecules.

POLARIZATION MEASUREMENTS ON CARBOXYLIC ACIDS IN DILUTE SOLUTION IN NON-POLAR SOLVENTS

BY HERBERT A. POHL,* MARCUS E. HOBBS, AND PAUL M. GROSS

From the Department of Chemistry, Duke University, Durham, North Carolina

The effective use of dielectric polarization measurements in investigating the association of polar molecules in non-polar solvents has been largely confined to a study of the association of alcohols. These show significant changes in their state of aggregation in a range of concentrations sufficiently high so that the equipment and methods hitherto available for dielectric constant measurements in solutions can be successfully applied to the investigation of their extent of association.

The case of the carboxylic acids in non-polar solvents is quite different. Freezing point determinations show that extremely low concentrations must be attained before significant changes in the state of aggregation occur. This range of concentrations in the case of many of the acids lies below 2×10^{-3} mole fraction. As a consequence it has been difficult to make sufficiently accurate measurements by most of the methods for molecular weight determination in solution to yield quantitative information about the molecular state of these acids in dilute solution in non-polar solvents. While it has been possible to use the freezing point method, this is applicable only at a single temperature and so gives information about temperature coefficients and energies of association only through the use of estimates made at other temperatures by other methods.

In this paper a description will be given of some of the improvements which were made in apparatus and methods in order to carry out dielectric polarization measurements in very low concentrations in non-polar solvents under anhydrous conditions. In view of the low concentrations involved it is necessary to include a comprehensive statement of the errors of the method. A detailed discussion of the methods employed for interpreting and analyzing the results from the standpoint of the association of the solutes is also included.

* Part of a thesis by Herbert A. Pohl submitted in partial fulfilment of the requirements for the degree of Ph.D. in Chemistry at Duke University.

THE CALCULATION OF POLARIZATION IN SOLUTIONS

The Debye equation for the determination of polarizations in solutions serves as a basis for the discussion of the modifications in experimental methods and procedure.

$$P_2 = \frac{M_1 f_1 + M_2 f_2}{f_2 d_{12}} \cdot \frac{E_{12} - 1}{E_{12} + 2} - \frac{f_1}{f_2} \cdot \frac{E_1 - 1}{E_1 + 2} \cdot \frac{M_1}{d_1} \quad (1)$$

The subscripts ₁ and ₂ and ₁₂ refer to solvent, solute and the solution respectively. P is the polarization, E the dielectric constant, M the molecular weight, d the density and f the mole fraction. Since in the dilute solutions under consideration the experimental problem is essentially one of precise determination of increments in E and d due to increasing additions of solute it is convenient to rearrange equation (1) by use of the relations

$$\frac{1}{f_2} - 1 = \frac{W_1}{W_2} \cdot \frac{M_2}{M_1}, \quad \Delta E = E_{12} - E_1 \text{ and } \Delta d = d_{12} - d_1$$

where W_1 and W_2 refer to the weights of solvent and solute. On putting $E = E_1$ and $d = d_1$, (1) becomes

$$P_2 = M_2 \frac{W_1}{W_2} \left[\frac{E + \Delta E - 1}{E + \Delta E + 2} \frac{1}{(d + \Delta d)} - \frac{E - 1}{E + 2} \frac{1}{d} \right] + M_2 \frac{E + \Delta E - 1}{E + \Delta E + 2} \frac{1}{d + \Delta d} \quad (2)$$

As the difference expression in the first term is the chief source of difficulties in computation it is advantageous to expand (2) giving

$$P_2 = \frac{\frac{W_1}{W_2} (K_1 \Delta E - K_2 \Delta d - K_3 \cdot \Delta E \cdot \Delta d) + M_2 (E + \Delta E - 1)}{(E + \Delta E + 2)(d + \Delta d)} \quad (3)$$

$$\text{where } K_1 = \frac{1}{E + 2}, \quad K_2 = \frac{E - 1}{d}, \quad K_3 = \frac{E - 1}{E + 2} \cdot \frac{1}{d}.$$

Equation (3) is an exact form of the Debye equation. It allows a very rapid computation especially if the specific rather than the molar polarization is calculated. For purposes of approximate calculation the term containing $\Delta E \cdot \Delta d$ may be dropped at concentrations less than 0.02 mole fraction with negligible error in most cases. The changes in the values of $(E + \Delta E + 2)$, $(E + \Delta E - 1)$ and $(d + \Delta d)$ produced by variation of ΔE and Δd are small and readily

made by inspection. As ΔE and Δd approach zero, equation (3) becomes the same as the Hedestrond¹ equation which has the form

$$P_2 = \frac{3 P_1}{E^2 + E - 2} \cdot \frac{\Delta E}{f_2} - \frac{P_1}{d} \cdot \frac{\Delta d}{f_2} + \frac{P_1}{M_1} \cdot M_2 \quad (5)$$

where

$$P_1 = \frac{E - 1}{E + 2} \frac{M_1}{d}$$

This equation, which assumes that $\frac{\Delta E}{f_2}$ and $\frac{\Delta d}{f_2}$ are constant for small values of f_2 , is useful for rapid approximation of P_2 from the observed values of $\frac{\Delta E}{f_2}$ and $\frac{\Delta d}{f_2}$ at any particular concentration. It is also a useful form to employ in discussing the magnitude of various errors as will be seen later.

For associated solutes which exist in solution in several different molecular states the value P_2 in the Debye equation is the average polarization per monomeric formula weight of the solute if the mole fraction f_2 of solute is calculated on the basis of the formula of the monomeric form and M_2 is taken as the molecular weight of the monomer.

EXPERIMENTAL EQUIPMENT

1. The oscillators and electrical circuits have been described previously.² In the present investigation they are used in conjunction with a liquid instead of a gas condenser.

2. The liquid condenser consists essentially of three concentric cylinders of monel, the second cylinder being the high potential one. The capacity was approximately 450 mmfd. and the liquid volume approximately 400 cc. The completed condenser was made so that it could be submerged completely. Two small tubes leading down to the condenser served to admit and withdraw the contents. The whole condenser was tubular in shape and permitted a flow of the thermostat water inside and out thus insuring a rapid and effective temperature control. FIGURE 1 show the details of the condenser. The high potential cylinder was insulated and supported by six quartz rods, two of which are shown in the figure. A device for switching from the liquid condenser to a 2000 mmfd. reference condenser has been previously described.²

¹ Hedestrond, G. Zeit. physikal. Chem. B 2: 428. 1929.

² Hobbs, M. E., Jacokes, J. W., & Gross, P. M. Rev. Sci. Instruments April, 1940.

3. The high potential leads were constructed as shown in FIGURE 2. Polystyrene discs may be substituted for the quartz without loss of insulating value.

4. All transfers of solvents, solute, and solutions were under anhydrous conditions insofar as possible. A "dry box" made from a

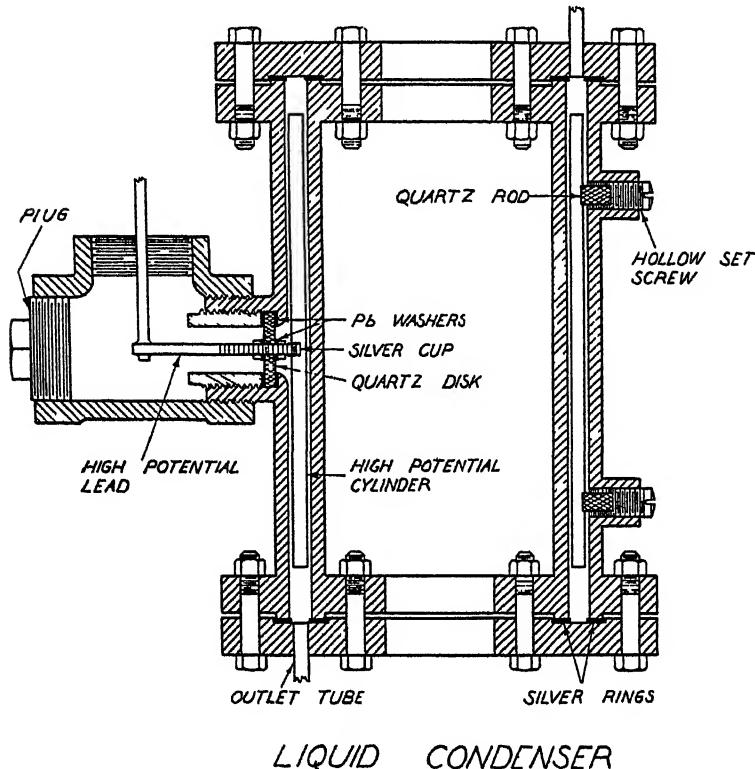


FIGURE 1.

section of an alcohol drum, greatly facilitated carrying out the transfers (FIGURE 2). CaCl_2 on the floor of the box was used as the desiccating agent. The transfer of solvent to the condenser was made by distilling directly into a transfer flask (FIGURE 3) and forcing the liquid from the flask into the condenser with dry nitrogen.

5. Stock solutions of the solute were introduced into the condenser by means of a weight burette (FIGURE 4). The stopcock was lubricated with "glydag." This lubricant was found to be quite satisfactory.

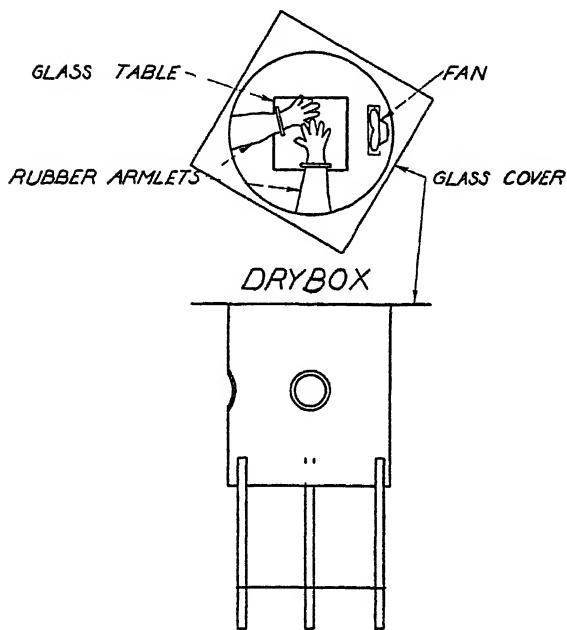
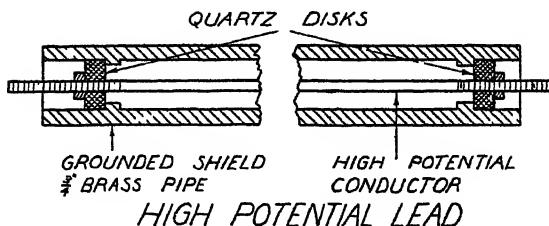


FIGURE 2.

6. The assembled equipment including the receiving flask (FIGURE 5) is shown in FIGURE 6. All outlets were protected from atmospheric moisture and only dry N_2 was allowed to enter the condenser during the period of a measurement.

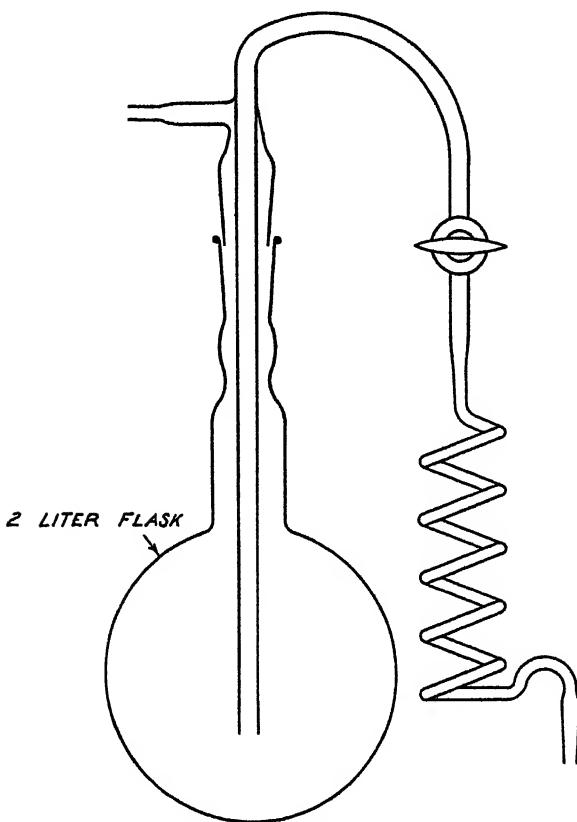
7. The pycnometers used in the density determinations were of the type described by Connell, Vosburgh, and Butler.³

8. The water thermostat used maintained an average temperature constant to $\pm 0.003^\circ C$ over a period of one day, and did not vary

³ Connell, L. C., Vosburgh, W. C., & Butler, J. A. V. Jour. Chem Soc 933 1938.

more than $\pm 0.03^\circ \text{ C}$ during two months of operation. The usual vacuum tube type control was employed.

9. The only calibration necessary was that of the capacity of the



TRANSFER FLASK

FIGURE 3.

liquid condenser. This was done with benzene and dry air assuming a value of 2.2627 for the dielectric constant of benzene at 30° C and a value of 1.00057 for that of air at 30° C and 1 atmosphere. The other parts of the apparatus had been previously calibrated.²

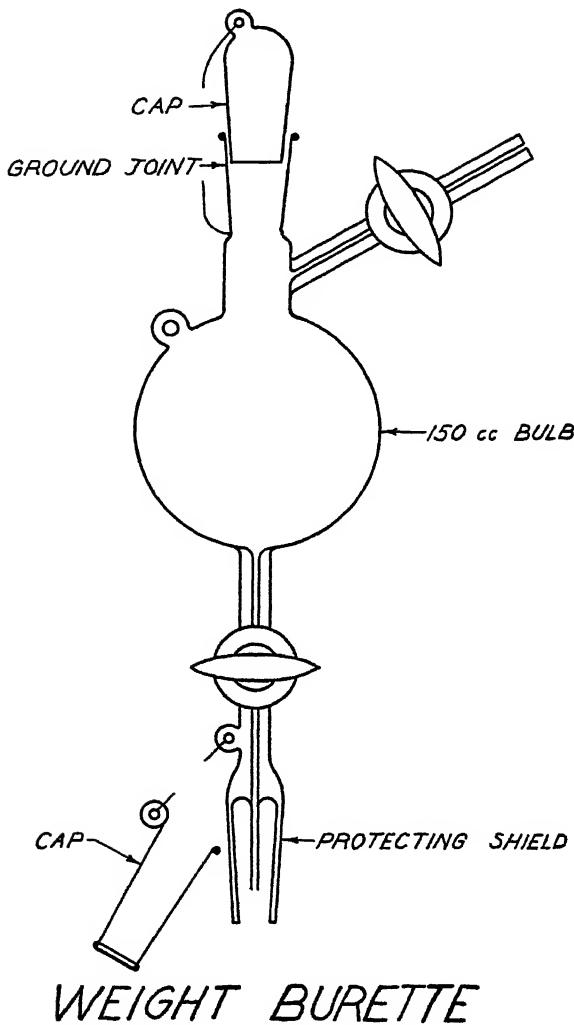
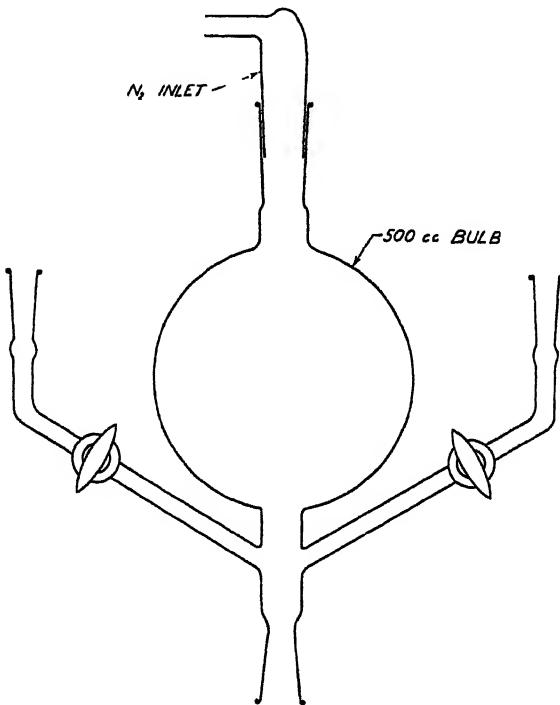


FIGURE 4.

DETAILS OF MEASUREMENTS

Prior to starting a measurement all equipment was carefully cleaned, dried and then rinsed with pure dry dust-free solvent. All equipment except the liquid condenser was then dried with dust-free air. The liquid condenser was further rinsed several times with pure dry solvent and then finally dried with a stream of N_2 which had passed over P_2O_5 .

A portion of the solvent present in the transfer flask was removed to be used in making the stock solution of solute. The condenser was then filled with some of the remaining solvent from the transfer flask and rinsed. Then it was filled again with solvent and a sample of solvent taken from the condenser for the density measurement. The volume of solvent in the condenser was accurately adjusted



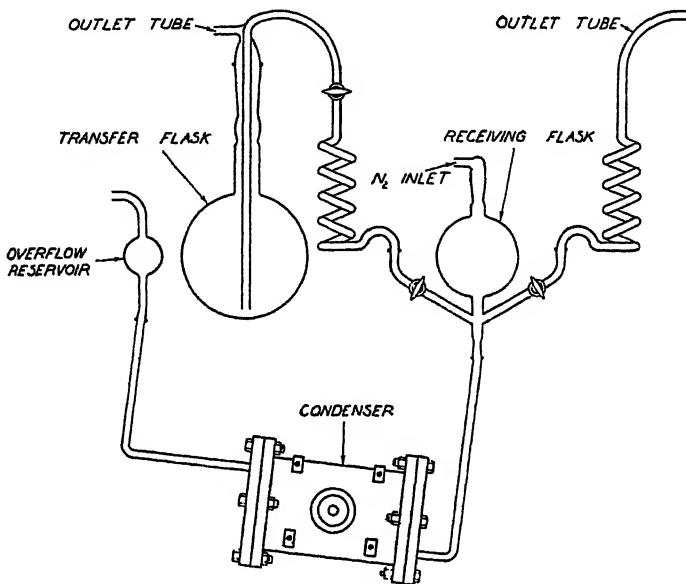
RECEIVING FLASK

FIGURE 5.

at this time. The solvent was then displaced several times from the condenser to the receiving flask by the pressure of dry N₂. The value of the capacitance of the condenser was found to shift very slightly during these displacements as will be noted in the section on errors.

After the value of the capacity of the solvent filled condenser had been determined a small quantity of stock solute solution was introduced into the receiving flask. The solvent and solute were

then mixed by two displacements of the liquid from the condenser into the receiving flask. This was shown to be adequate. Fifteen minutes after the mixing the solution had come to constant temperature so the capacity increment, ΔC , in the capacity of the condenser could be measured. Readings of capacity were always referred to a reference capacity so that drifts in the oscillator frequencies were practically eliminated. Successive additions of solute were made in



ASSEMBLED APPARATUS

FIGURE 6.

the manner described above and ΔC measured in each case. The density of the final solution was determined in the same pycnometer used for determining the density of the solvent, thus the total density, increment, Δd , was determined rather precisely. Δd is given by $\frac{\Delta M}{V}$ where ΔM is the difference in mass of the pycnometer when filled with solvent and then filled with solution. V is the pycnometer volume. The values of ΔE , or dielectric constant increments for the various solutions were calculated by $\frac{\Delta C}{C_{vac}}$ where C_{vac} is the vacuum capacity of the liquid condenser.

All measurements were made at $30.00 \pm .05^\circ$ C. The variation of the condenser temperature during a particular measurement did not exceed $\pm 0.002^\circ$ C.

PREPARATION AND PURIFICATION OF MATERIALS

Benzene

Kahlbaum benzene which had been dried over P_2O_5 was refluxed over sodium for at least ten hours and then distilled through a 180 cm. Dufton column. All except a small low-boiling fraction distilled within 0.01° range. The benzene was further purified by fractional crystallization until a freezing range of 0.01° C or less was obtained. Before use in a measurement, the benzene was refluxed over sodium for at least two hours and then after two hundred cc. were distilled to rinse moisture from the distilling column, the remainder was distilled directly into the transfer flask.

The boiling point found was 80.10° , $d_{4^\circ}^{80^\circ} = 0.86835$. Timmermans⁴ gives B. P. = 80.20° , $d_{4^\circ}^{80^\circ} = 0.86844$. Wojciechowski⁵ gives B. P. = $80.093^\circ \pm 0.002^\circ$ on benzene from four sources. Cohen and Buij⁶ give B. P. = 80.2° and $d_{4^\circ}^{80^\circ} = 0.86844$.

Heptane

The heptane used had been previously washed with sulphuric acid, sodium hydroxide and then fractionally distilled from sodium. This material was twice redistilled from sodium before use in the present work. It boiled over a 0.04° C range. This was refluxed for six hours over sodium before distilling it into the transfer flask for the measurement. $d_{4^\circ}^{80^\circ} = 0.67475$.

Acetic Acid

Two liters of C. P. acid was found to freeze at 16.0° . This was fractionally crystallized until the freezing point was 16.4 , and was then fractionally distilled in a Widmer column. The freezing point of $16.63^\circ \pm 0.05^\circ$ was unchanged by further crystallization. Timmermans⁷ gives 16.55° C as the freezing point.

Formic Acid

A sample of Kahlbaum formic acid was fractionally crystallized once and then distilled in an all-glass Widmer still. B. P. 100.2° ,

⁴ Timmermans, J., & Martin, F. *Jour. Chim. Phys.* 23: 733. 1926.

⁵ Wojciechowski, *Roczniki Chem.* 16: 534. 1936.

⁶ Cohen, E., & Buij, J. S. *Zeit. Physikal. Chem.* B 35: 270. 1937.

⁷ Timmermans, M. J., & Hennaut-Roland, Mme. *Jour. Chim. Phys.* 27: 401. 1930.

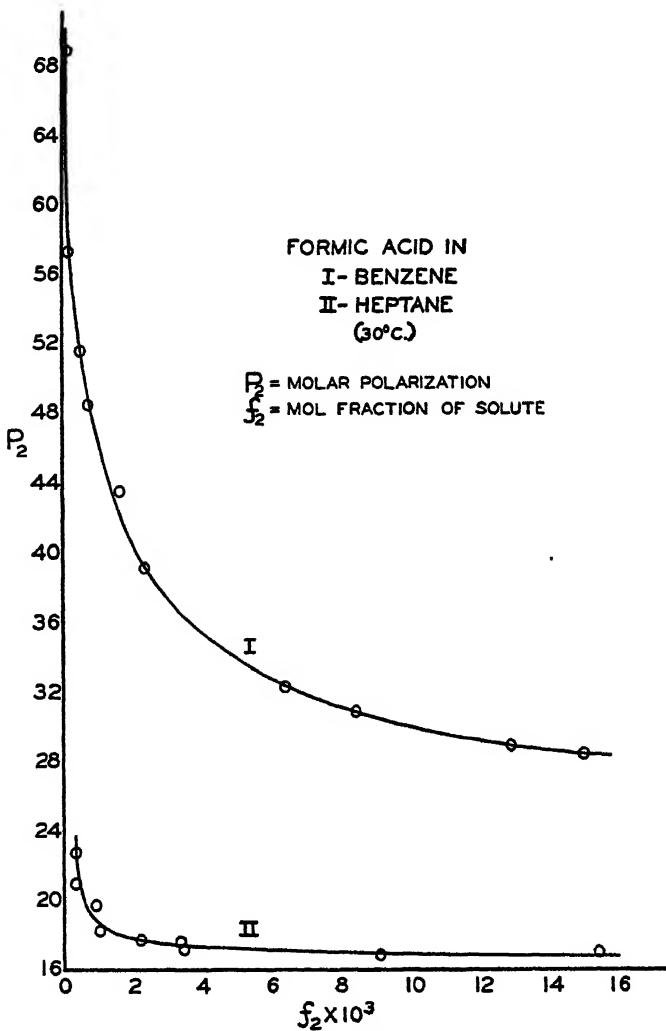


FIGURE 7.

uncorr., range 0.15° C. This acid was fractionally crystallized six times and was found to melt at 8.5° C \pm 0.1 over a range of 0.07° C. Timmermans⁷ gives 8.40° C. The critical solution temperature in anhydrous benzene according to the method of Ewins⁸ was determined and found to be $73.4^\circ \pm 0.3^\circ$ C. thus indicating that the present sample probably contains less than 0.02% water.

⁸ Ewins, A. J. Jour. Chem. Soc. 105: 530. 1914.

Propionic Acid

Fractional distillation is regarded by Timmermans⁷ as a satisfactory method of purification. A sample of Kahlbaum acid was fractionally distilled in a Widmer column four times. The final sample of 45 grams boiled over a range of 0.07° C. B. P. = $141.5 \pm 0.2^{\circ}$;

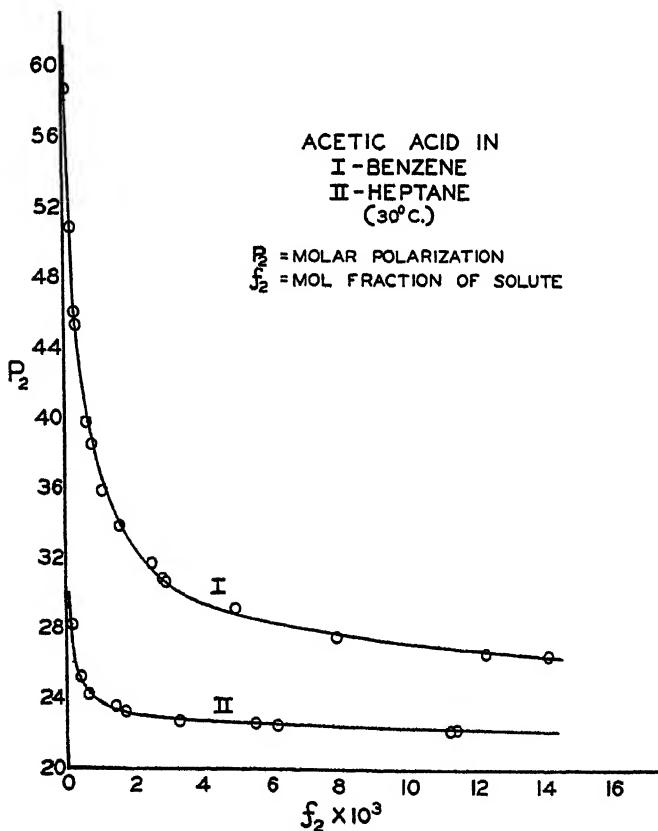


FIGURE 8.

Timmermans⁷ gives 141.35° C. A rough determination of the freezing point gave a value of -22° C $\pm 1^{\circ}$ C. Timmermans gives -20.8° C.

Trimethyl Acetic Acid

A sample of Eastman Kodak material was fractionally distilled, B. P. $164.15 \pm 0.20^{\circ}$ C. After distillation it was fractionally crystallized three times, redistilled and the crystals dried over P_2O_5 for twenty-four hours.

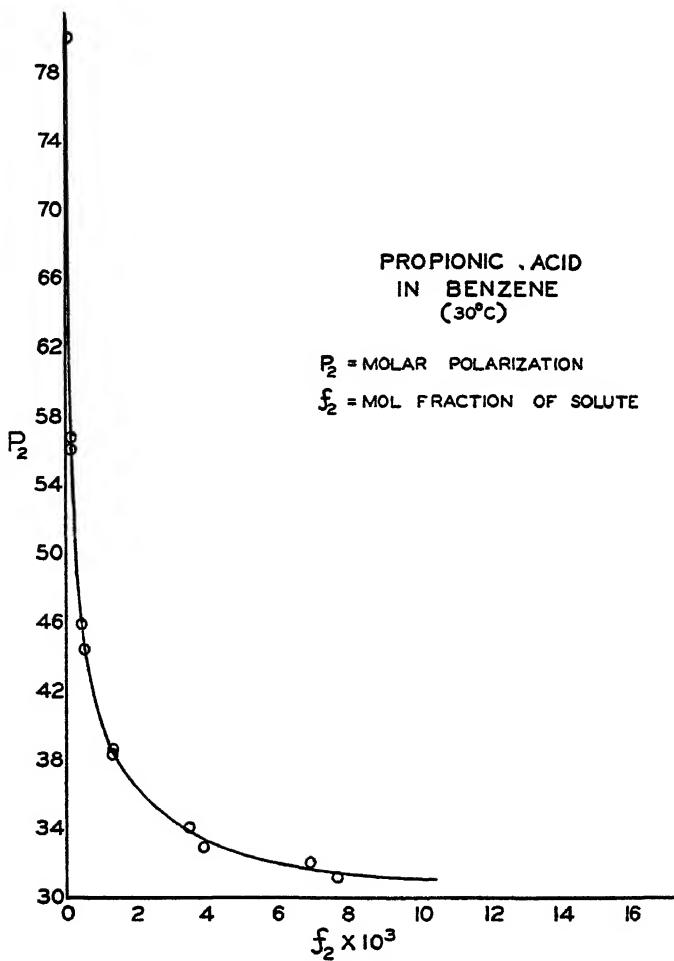


FIGURE 9.

Monochloro Acetic Acid

An Eastman Kodak sample was fractionally crystallized three times and was then recrystallized six times from benzene. The crystals after each precipitation were washed with cold benzene and dried over 99% H_2SO_4 in a vacuum desiccator in accordance with the method suggested by P. M. Gross.⁹ M. P. = $62.3 \pm 0.5^\circ$, range = 0.2° .

⁹ Gross, P. M. Ph.D. Dissertation, Columbia University. 1919.

Benzoic Acid

A sample of standard benzoic acid from the Bureau of Standards was used without further treatment.

Butyric Acid

Kahlbaum acid was fractionally distilled three times in a Widmer column and then fractionally crystallized twice. The product of the second crystallization gave a melting point of -5.7° C with a range of 0.1° C.

EXPLANATION OF TABLES

The following tables contain the data obtained by measurement of the dielectric constants, densities and mole fractions of the various solutions.

f_2 = mole fraction of the solute

ΔE = increment in the dielectric constant of the solution over that of the solvent

P_2 = average polarization of the solute per molecular weight M_2 of solute stated

$\frac{\Delta d}{f_2}$ = ratio of the increment of density to mole fraction.

For convenience of reference the constants for the solvents which were used in calculating the polarizations are listed below.

$M_{Benzene} = 78.05$

$M_{Heptane} = 100.19$

d_4^{30} Benzene = .86835

d_4^{30} Heptane = .67475

$E_{Benzene}$ (30°) = 2.2627

$E_{Heptane}$ (30°) = 1.910

The last value is an average of the value given by Smyth¹⁰ and that of deBruyne.¹¹

¹⁰ Smyth, G. P., & Walls, W. S. Jour. Am. Chem. Soc. 54: 1859. 1932.

¹¹ deBruyne, J. M. A., Davis, R. M., & Gross, P. M. Jour. Am. Chem. Soc. 55: 3936. 1933.

TABLE 1
CHLOROBENZENE IN BENZENE (30° C)

$f_2 \times 10^3$	$\Delta E \times 10^3$	P_2
0.0683	0.236	81.6
0.2408	0.836	81.8
0.3408	1.189	82.1
0.3730	1.284	81.4
0.5192	1.810	82.1
0.6606	2.317	82.4
0.7756	2.710	82.2
0.8690	3.020	81.9
1.089	3.829	82.5
2.140	7.517	82.4
2.642	9.280	82.5
5.810	20.470	82.6
8.300	28.870	82.0

$$\frac{\Delta d}{f_2} = 0.264 \quad P_2^\infty = 82.2 \quad \mu = 1.58 \times 10^{-18}$$

$$M_2 = 112.56 \quad R_D = 31.4$$

TABLE 2
FORMIC ACID IN BENZENE (30° C)

$f_2 \times 10^3$	$\Delta E \times 10^3$	P_2
0.1121	0.423	68.8
0.1601	0.479	57.2
0.4902	1.280	51.5
0.6981	1.673	48.4
1.634	3.161	41.5
2.286	4.045	39.0
6.392	8.376	32.2
8.464	10.294	30.7
12.965	14.129	28.8
15.103	15.931	28.3

$$\frac{\Delta d}{f_2} = 0.0946 \quad M_2 = 46.03$$

TABLE 3
FORMIC ACID IN HEPTANE (30° C)

$f_2 \times 10^3$	$\Delta E \times 10^3$	P_2
0.2877	0.091	20.9
0.3235	0.123	22.8
0.8806	0.240	19.6
0.9900	0.218	18.1
2.233	0.454	17.6
3.368	0.677	17.5
3.459	0.639	17.1
9.162	1.572	16.7
15.515	2.812	17.0

$$\frac{\Delta d}{f_2} = 0.0820 \quad M_2 = 46.03$$

TABLE 4
ACETIC ACID IN BENZENE (30° C)

Series	$f_2 \times 10^3$	$\Delta E \times 10^3$	P_2
(1)	0.0548	0.150	58.7
(2)	0.1351	0.299	50.9
(1)	0.2506	0.475	46.0
(2)	0.2885	0.531	45.3
(2)	0.5894	0.865	39.8
(3)	0.7380	1.017	38.5
(1)	1.039	1.249	35.8
(4)	1.541	1.634	33.7
(2)	1.557	1.661	33.8
(3)	2.498	2.315	31.7
(1)	2.826	2.454	30.9
(4)	2.931	2.502	30.6
(3)	4.940	3.709	29.1
(1)	7.946	5.146	27.6
(1)	12.37	7.236	26.6
(3)	14.18	8.323	26.5

$$\frac{\Delta d}{f_2} = 0.0817 \quad M_2 = 60.05$$

TABLE 5
ACETIC ACID IN HEPTANE (30° C)

$f_2 \times 10^3$	$\Delta E \times 10^3$	P_2
0.1543	0.072	28.2
0.3771	0.136	25.2
0.6091	0.198	24.2
1.3903	0.418	23.5
1.7120	0.499	23.2
3.3053	0.911	22.8
5.544	1.497	22.6
6.218	1.650	22.5
11.302	2.917	22.2
11.531	2.998	22.3
$\frac{\Delta d}{f_2} = 0.1164$		$M_2 = 60.05$

TABLE 6
PROPIONIC ACID IN BENZENE (30° C)

$f_2 \times 10^3$	$\Delta E \times 10^3$	P_2
0.0191	0.073	80.0
0.1037	0.237	56.7
0.1112	0.249	56.1
0.4248	0.660	45.9
0.4780	0.694	44.4
1.277	1.333	38.3
1.302	1.371	38.5
3.474	2.616	34.0
3.883	2.631	32.9
6.908	4.245	32.0
7.695	4.281	31.1
$\frac{\Delta d}{f_2} = 0.0781$		$M_2 = 74.05$

TABLE 7
BUTYRIC ACID IN BENZENE (30° C)

$f_2 \times 10^3$	$\Delta E \times 10^3$	P_2
0.0569	0.188	76.4
0.0842	0.193	61.2
0.1493	0.331	60.2
0.3080	0.508	51.8
0.5408	0.856	50.8
0.6787	0.923	47.4
0.8907	1.221	47.6
1.063	1.198	43.8
1.517	1.588	42.8
3.223	2.534	38.9
5.627	3.977	37.7
8.510	5.000	36.0
10.330	5.798	35.6
10.565	6.428	36.3

$$\frac{\Delta d}{f_2} = 0.0906 \quad M_2 = 88.06$$

TABLE 8
TRIMETHYL ACETIC ACID IN BENZENE (30° C)

$f_2 \times 10^3$	$\Delta E \times 10^3$	P_2
0.0376	0.090	72.4
0.1082	0.205	61.8
0.4162	0.482	52.4
0.6252	0.604	47.9
1.118	0.899	46.6
2.102	1.411	43.6
3.247	1.746	42.3
3.764	2.027	41.6
8.235	3.195	39.9

$$\frac{\Delta d}{f_2} = 0.0423 \quad M_2 = 102.13$$

TABLE 9
BENZOIC ACID IN BENZENE (30° C)

$f_2 \times 10^3$	$\Delta E \times 10^3$	P_2
0.1154	0.361	77.7
0.1627	0.465	73.7
0.3516	0.856	67.5
0.5825	1.266	63.6
1.342	2.473	58.7
2.368	3.919	56.9
4.357	6.528	53.6
5.482	7.967	52.9
8.096	11.134	51.7
9.727	12.150	51.4

$$\frac{\Delta d}{f_2} = 0.3357 \quad M_2 = 122.12$$

TABLE 10
MONOCHLOROACETIC ACID IN BENZENE (30° C)

$f_2 \times 10^3$	$\Delta E \times 10^3$	P_2
0.1051	0.665	114.3
0.1358	0.841	112.2
0.6193	3.120	95.1
0.6328	3.189	95.2
1.862	7.901	83.3
1.884	7.970	83.1
4.729	17.516	75.1
5.248	19.167	74.4
10.066	33.984	70.1
11.095	37.196	69.7

$$\frac{\Delta d}{f_2} = 0.3846 \quad M_2 = 94.70$$

DISCUSSION OF ERRORS

As will be indicated later the present measurements extend to a range of concentrations considerably lower than the measurements of most previous workers. It is desirable therefore to discuss in some detail, at this point, the sources and magnitudes of the errors which may affect the measurements.

The Hedstrand approximation formula¹ for the molar polarization given by

$$P_2 = P_1 \frac{M_2}{M_1} + P_1 \frac{3(\Delta E)}{(E^2 + E - 2)f_2} - P_1 \frac{\Delta d}{f_2 d}$$

is well adapted to the discussion of errors in dilute solution polarization measurements.

Most of the errors in P_2 may be conveniently considered from the standpoint of separate errors in the quantities E , M_1 , M_2 , d , f_2 , and ΔE , and Δd . As may be anticipated errors in ΔE will be the ones which seriously limit the precision obtainable in the dilute solution region. The effect of an uncertainty in E , M_1 , M_2 , and d may be shown¹² to be given by

$$\frac{\delta P_2}{P_2} = \left(1 - \frac{\Delta d}{d \cdot f_2} \frac{P_1}{P_2} \right) \frac{\delta d}{d} \quad (1)$$

$$\frac{\delta P_2}{P_2} = \left(3 + \frac{P_1}{P_2} (2E + 1) \frac{\Delta E}{f_2} \right) \frac{E \cdot \delta E}{(E - 1)(E + 2) \cdot E} \quad (2)$$

$$\frac{\delta P_2}{P_2} = \frac{P_1 M_2}{P_2 M_1} \cdot \frac{\delta M_2}{M_2} \quad (3)$$

$$\frac{\delta P_2}{P_2} = \left[\frac{3 \Delta E}{(E^2 + E - 2) \cdot f_2} - \frac{\Delta d}{f_2 \cdot d} \right] \frac{P_1}{P_2} \frac{\delta M_1}{M_1} \quad (4)$$

Similarly the effect of an uncertainty in ΔE , Δd , and f_2 is given by

$$\frac{\delta P_2}{P_2} = \frac{P_1 \cdot 3 \cdot \Delta E}{P_2 \cdot (E^2 + E - 2) \cdot f_2} \frac{\delta(\Delta E)}{\Delta E} \quad (5)$$

$$\frac{\delta P_2}{P_2} = \frac{P_1 \cdot \Delta d}{P_2 \cdot d \cdot f_2} \frac{\delta(\Delta d)}{\Delta d} \quad (6)$$

$$\frac{\delta P_2}{P_2} = \left[\frac{P_1 \cdot \Delta d}{P_2 \cdot d \cdot f_2} - \frac{P_1 \cdot 3 \cdot \Delta E}{P_2 (E^2 + E - 2) \cdot f_2} \right] \frac{\delta f_2}{f_2} \quad (7)$$

In general the values of the variables are such as to make the fractional error in P_2 approximately equal to the fractional errors of d in

¹- Muller, F. H. Physikal. Zeit. 33: 689; Trans. Faraday Soc. 30: 729. 1930.

(1), of M_2 in (3), of M_1 in (4) and of ΔE in (5). The fractional errors in P_2 are less than the fractional error in Δd in (6) and about equal to the fractional error of f_2 in (7). In certain cases the frac-

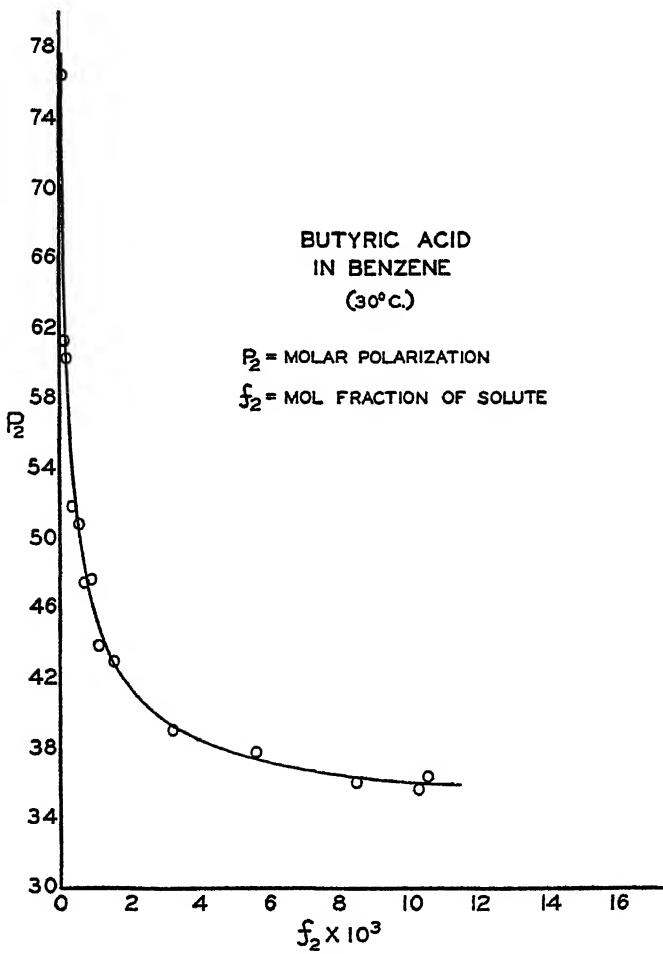


FIGURE 10.

tional error of P_2 may become as much as eight times the fractional error in E in (2), however, the factor will usually not exceed four. The estimated errors in parts per thousand for d , E , M_2 , and M_1 are respectively 0.1, 0.3, 1.0, and 1.0. Thus we may in general expect an error in the absolute magnitude of P_2 , from these sources of the order of about 4 parts per thousand, or 0.4%. The errors in ΔE are, in

the present experiments, the most important, and will be given detailed consideration. ΔE is defined experimentally by

$$\Delta E = \frac{\Delta c_{ap.}}{C_{vac}}$$

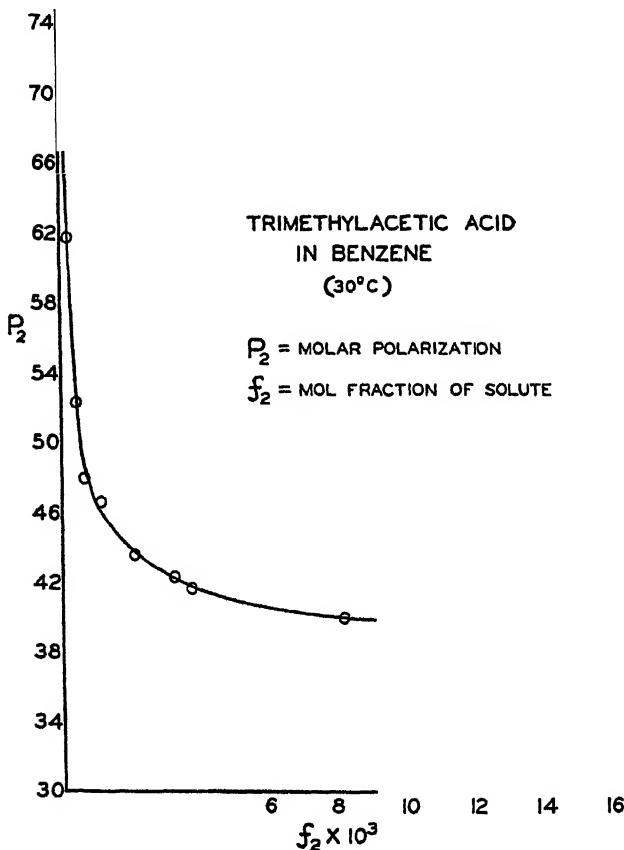


FIGURE 11.

where $\Delta c_{ap.}$ is the change in capacity of the liquid condenser and C_{vac} is that part of the vacuum capacity of the liquid condenser which is available for change by means of introduction of a dielectric medium. Since C_{vac} may be rather easily established to a precision of 0.2—0.4 parts per 1000, we may expect the errors in ΔE to originate almost entirely from errors in $\Delta c_{ap.}$.

The chief errors in Δ_{Cap} . may be listed as the following. (1) Errors due to the inductance of the leads between the measuring condenser and the liquid condenser. (2) The effect of the resistance of the solution being measured on the effective capacitance of C_{liq} . (3) Change of the reference air condenser capacitance during a measurement. This may be caused by a change in pressure, or by a change in temperature of the air. (4) Change of the temperature of the liquid in the liquid condenser during a measurement. (5) A

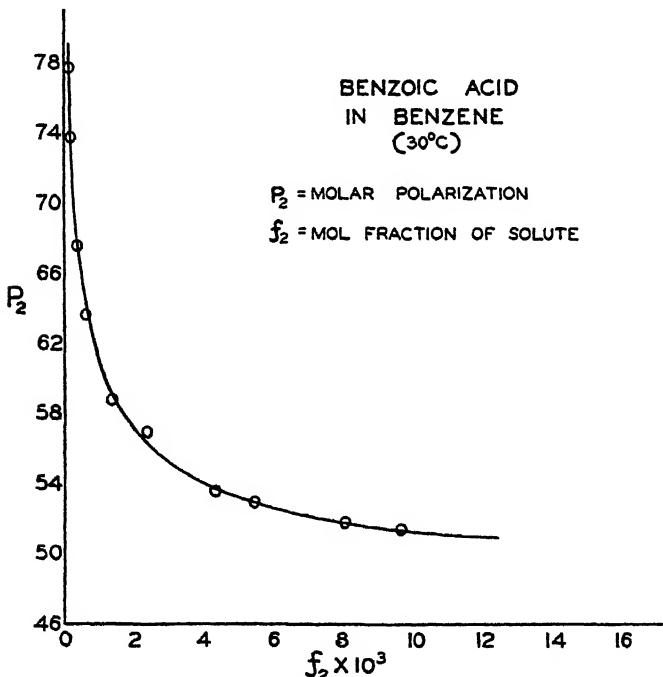


FIGURE 12.

change in frequency of the oscillators during the course of a measurement of the reference capacitance and the liquid capacitance. (6) A change which appears as an apparent change in the E of the liquid in the course of repeated displacements of the liquid from, and its return to, the condenser even though no solute is added. (7) The uncertainty in setting and reading the position of the meter arm on the calibrated glass scale. (8) A calibration uncertainty of approximately 1 part per 1000. There may be other small error contributions such as failure of a linear relation between measured capacity differences of the liquid condenser and computed dielectric values

even in the small range of dielectric values involved. It is felt, however, that the lumped effect of all these would not significantly change the estimates which follow.

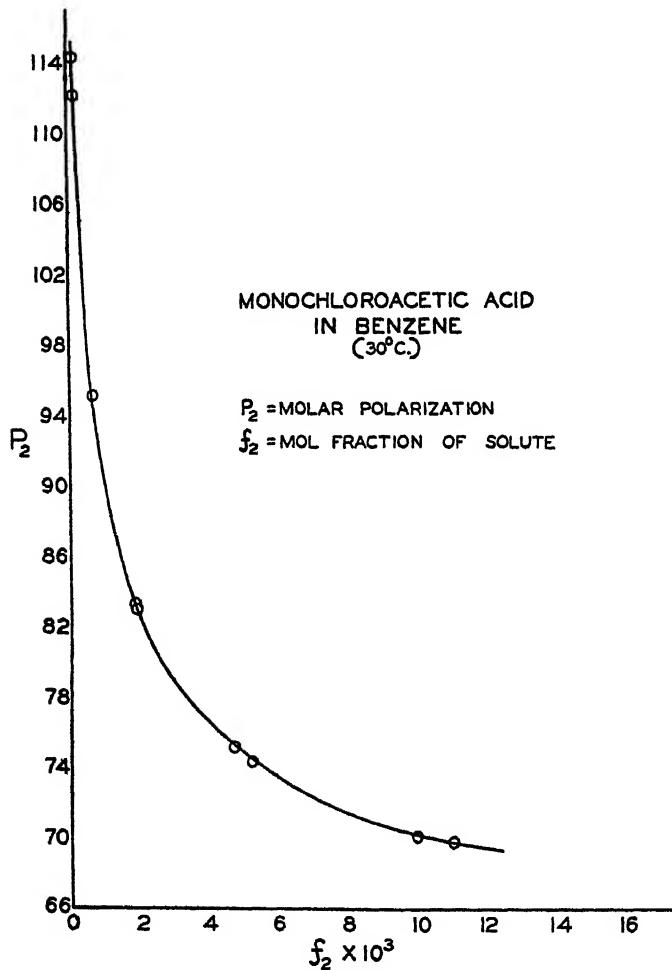


FIGURE 18.

Considering these uncertainties in the order named and without regard to sign we find that the inductance correction may be made relatively small if the value of C_{vac} is determined through the same lead system as Δ_{gap} . The remaining error is of the order of 3-5 parts per 1000 in the absolute value of ΔE .

The resistance of the solution changes with each addition of solute, but as may be seen from the following equation the changes produced are very small if C , R and W are large.

$$C^1 = \frac{1 + W^2 C^2 R^2}{W^2 C^2 R^2} \cdot C$$

Here C^1 is the effective capacitance of C if it is in a circuit where $W = 2\pi f$, the frequency being f , and R is the resistance considered as a shunt across the condenser. Estimation of R offers some difficulty but if one assumes as an approximation that the conductivity of acetic acid solutions in benzene is proportional to the mole fraction and uses the value of P. M. Gross⁹ for the specific conductivity of pure acetic acid of 2.5×10^{-8} ohms⁻¹, the conductivity may be calculated by

$$c_{1,2} = f_1 c_1 + f_2 c_2$$

where $c_{1,2}$, c_1 and c_2 are conductivities of the solution, the solvent, and the solute, and f_1 and f_2 are mole fractions. Using $c_1 = 3 \times 10^{-15}$ one obtains

$$R = 1.5 \times 10^8 \text{ ohms.}$$

Since $W = 3 \times 10^6$ and $C = 5 \times 10^{-10}$ farads, one obtains

$$C^1 = \frac{1 + 5 \times 10^{10}}{5 \times 10^{10}} C$$

In this case the effective capacity is practically identical with the real capacity and the resistance factor is negligible.

The change of the reference air capacitance due to pressure may be calculated in the following way

$$\delta E = \frac{3P}{RT} \delta p \simeq 8 \times 10^{-7} \delta p$$

where P = molar polarization of air, and p is pressure, R the gas constant, and T the absolute temperature. If $\delta p = 0.2$ mm.

$$\delta E \simeq 0.2 \times 10^{-6}$$

and if the δp occurred during a measurement of $\Delta_{O_{ap}}$, the error in ΔE is $(\delta E) \frac{E_a}{E_a}$ and is then $\delta(\Delta E) \simeq (0.2 \times 10^{-6}) (2) \simeq 0.4 \times 10^{-6}$.

The error due to change of temperature is readily seen to be

$$\delta E = -\frac{3P}{RT^2} \delta T \simeq 2 \times 10^{-6} \delta T$$

and since the reference condenser is thermostated we may assume

$$\delta T = 0.1^\circ\text{C}.$$

$$\delta E \simeq 0.2 \times 10^{-6}$$

$$\delta(\Delta E) \simeq (0.2 \times 10^{-6}) (2) \simeq 0.4 \times 10^{-6}$$

The effect of the change of temperature of the liquid on ΔE is readily obtained by the relation for benzene

$$\delta E = -0.002 \delta T$$

Since the liquid condenser was thermostated so that maximum expected change during a measurement of ΔC_{op} . was less than 0.002°C we may write

$$\delta(\Delta E) = 4 \times 10^{-6}$$

The uncertainty in ΔE due to a change in the frequency, f , during the time needed for comparison of the liquid condenser capacitance with the reference condenser may be calculated in the following manner

$$\delta C = -2 \frac{C}{f} \delta f$$

and since $C \simeq 4700 \Delta S_1$ units and $f \simeq 5 \times 10^{-5}$ cycles/sec one obtains

$$\delta C \simeq 19 \times 10^{-3} \delta f$$

Since f was stable to better than 0.2 cycles per second for this interval, we may write

$$\delta C' \simeq 3.8 \times 10^{-3}$$

or since

$$\delta E = \frac{\delta C}{C_{vac}} \simeq \frac{3.8 \times 10^{-3}}{912} \simeq 4 \times 10^{-6}$$

or $\delta(\Delta E) \simeq (2)(4 \times 10^{-6}) = 8 \times 10^{-6}$

Since two E readings must be made to obtain ΔE .

The uncertainty in ΔE due to the apparent changes in E when the liquid is displaced from, and then returned to the condenser was found by repeated trial, to be an effect amounting on the average to approximately 7×10^{-6} in ΔE , therefore from this source

$$\delta(\Delta E) = 7 \times 10^{-6}$$

The uncertainty in reading the difference between the capacitance of the reference and of the liquid condenser is approximately 2×10^{-6} in E and since two such differences must be measured

$$\delta(\Delta E) \simeq 4 \times 10^{-6}$$

The calibration uncertainty of 1 in 10^8 for Δ_{Cap} leads to

$$\delta(\Delta E) = 1 \times 10^{-6}$$

In order to calculate the total variable uncertainty in ΔE without regard to sign we must sum all contributions except that contributed by the inductance. These are tabulated below.

Source	Error in $(\Delta E) \times 10^6$
Reference condenser	
P—variation.....	0.4
T—variation.....	0.4
Liquid condenser	
Temperature.....	4
Frequency changes.....	8
Displacement changes.....	7
Reading of arm positions.....	4
Calibration.....	1
	<hr/>
	25

This total $\delta(\Delta E)$ of 25×10^{-6} leads one to anticipate rather larger deviations in P_2 than are actually observed. This might well be the case as cancellation of errors of opposite sign occurs and no account has been taken of this in the summation. In general the observed deviations are of the order of 2–3 cc. in P_2 in solutions where $f_2 = 1 \times 10^{-4}$ and approximately 0.2–0.3 cc in P_2 where $f_2 = 1 \times 10^{-2}$.

The error in Δd is not particularly important as the entire Δd term in the Hedstrand formula is relatively small and seldom exceeds, in the present cases, 25% of the total P_2 value. Since Schulz¹³ has shown that the densities of several carboxylic acids in benzene, cyclohexane and dioxane vary linearly with concentrations over the range $f_2 = 5 \times 10^{-4}$ to 6×10^{-2} , the assumption that $\frac{\Delta d}{f_2}$ is constant was made in all calculations. The measurement of Δd involved the measurement of ΔM , the increase in mass of a pycnometer when filled with a solution over that when filled with solvent, and V , the volume of the pycnometer respectively, viz.

$$\Delta d = \frac{\Delta M}{V}$$

$$\frac{\delta(\Delta d)}{d} = \frac{\delta(\Delta M)}{M} + \frac{\delta(d)}{d} - \frac{\delta(V)}{V}$$

¹³ Schulz, G. Zeit. Physikal. Chem. B 40: 151. 1938.

Now δV is of order of 0.2 per 1000 therefore $\delta(\Delta d)$ is negligibly effected by an error in V . The actual error in ΔM from adjusting levels is approximately 0.4 mg., from temperature changes approximately 0.3 mg., and from weighing approximately 0.2 mg. Choosing an average value of $\Delta M = 100$ mg. one obtains a fractional error in Δd of 9 parts per 1000. In general the actual observed reproducibility of the Δd is about 4 parts per 1000.

The uncertainty of f_2 may be evaluated from the following relations.

$$\begin{aligned}\frac{1}{f_2} &= \frac{W_1 M_2}{W_2 M_1} + 1 \\ \frac{\delta f_2}{f_2} &= -f_2 \cdot \frac{M_2 W_1}{M_1 W_2} \cdot \frac{\delta W_1}{W_1} \\ \frac{\delta f_2}{f_2} &= f_2 \cdot \frac{M_2 W_1}{M_1 W_2} \cdot \frac{\delta W_2}{W_2}\end{aligned}$$

where W_1 is weight of solvent, W_2 weight of solute, and M_1 and M_2 the molecular weight of solvent and solute respectively. As may be readily seen the fractional error in f_2 is approximately equal to the fractional error in W_1 and in W_2 and f_2 is small compared to 1. Because of the very dilute solutions required it was found necessary to use stock solutions made up from the solute and solvent in suitable proportions. A detailed consideration, based on experimental measurements of losses during transfer and calculation of losses due to vaporization, leads to an anticipated average error in f_2 of about 1-2 parts per 1000 in very dilute solutions and about 0.5 part per 1000 in more concentrated regions.

A source of error not previously mentioned is that arising from impurities. The impurity may be in the solute, or in the solvent. Inspection of the freezing point, and boiling point ranges of the solutes used in this investigation indicates that, as a close approximation, the total impurities present did not exceed 0.5%. It is estimated that their water content in no case exceeded 0.1%.

The impurities in the solvent may be either of two kinds, (1) those that have no specific effects, or reaction on or with, the solute and (2) those which do react with the solute, *e. g.*, water. Reasonable purification of the solvent in regard to impurities of the first class is sufficient to reduce this source of error until it is negligible. For substances of the second class a much more careful purification and continued care in handling is necessary. This case may be illustrated by supposing an equi-mole fraction mixture of water and acid in

benzene as a solvent. As an estimate we might anticipate a 50% change in the P_2 which would be observed as compared with that which would have been observed if no water had been present. Thus, if we wish to measure compounds at concentrations where $f_2 = 10^{-4}$ with a precision of 1 part per 1000, impurities such as water would have to be reduced to less than 2×10^{-7} mole fraction. This estimate shows how important it is that reactive impurities be removed from the solvent.

In TABLE 11 the errors which effect the reproducibility of P_2 are

TABLE 11

ERRORS COMPUTED FOR THE CASE OF ACETIC ACID IN BENZENE

$\Delta E = 0.15 \times 10^{-4}$	$\Delta E = 0.40 \times 10^{-3}$	$\Delta E = 1.0 \times 10^{-3}$	$\Delta E = 5.0 \times 10^{-3}$	
$f_2 = 5 \times 10^{-6}$	$f_2 = 2 \times 10^{-4}$	$f_2 = 1 \times 10^{-3}$	$f_2 = 1 \times 10^{-2}$	
Source	$\delta(\Delta E) \times 10^{-6}$			
Reference Condenser				
P variation.....	0.4	0.4	0.4	0.4
T variation.....	0.4	0.8	1.2	1.6
Liquid Condenser				
Temperature.....	4	4	4	4
Frequency changes....	8	8	8	8
Displacement changes..	7	14	21	28
Reading of arm positions	4	4	4	4
Calibrations.....	1	1	1	1
Total.....	25	32	40	47
$\delta P_2 = \frac{P_1 \cdot 3 \cdot \delta(\Delta E)}{(E^2 + E - 2) \cdot f_2} \dots$	7.2 cc.	2.4 cc.	0.6 cc.	0.07 cc.
Errors in Δd		$\delta(\Delta d)$ in parts per 1000		
V	0.2	0.2	0.2	0.2
ΔM	9	9	9	9
Total.....	9	9	9	9
$\delta P_2 = \frac{P_1 \cdot \delta(\Delta d)}{d \cdot f_2} \dots$	0.03 cc.	0.03 cc.	0.03 cc.	0.03 cc.
Errors in f_2		δf_2 in parts per 1000		
Making up stock				
solutions.....	+0.3	+0.3	+0.3	+0.3
Additions of stock				
solution.....	-6.0	-4.0	-1.0	-0.2
Mixing losses.....	+0.4	+0.8	+1.2	+2.0
Total.....	5	3	0.5	2
$\delta P_2 = \text{equation (7).} \dots$	0.2 cc.	0.1 cc.	0.01 cc.	0.02 cc.
Errors due to impurity				
In solute $\approx (0.005) (60)$	0.3 cc.	0.3 cc.	0.3 cc.	0.3 cc.
Overall estimate of ex- pected reproducibility of P_2	8 cc.	2.8 cc.	0.9 cc.	0.4 cc.

applied to the particular case of solutions of acetic acid in benzene in different ranges of concentration.

The estimate above for the lowest f_2 values is definitely larger than the spread of our observed points, the latter being of the order of 2–3 cc. in the more dilute regions. As mentioned previously this difference is doubtless attributable to cancellation of errors of opposite sign, or possibly to an overestimate of the magnitude of some of the errors. The error in the absolute value of P_2 is probably of the order of 0.5 cc. to 1.0 cc. for the more concentrated and most dilute points respectively.

DISCUSSION OF RESULTS

Certain of the results which bear on the experimental precision attained in relation to the work of others may be considered first. The upper part of FIGURE 14 shows a comparison of our results for chlorobenzene solutions in benzene with those of Tiganik¹⁴ and Müller.¹² It will be observed that the scattering of our points is little if any greater than theirs, although the lower limit of the concentrations which we have measured is an order of magnitude lower.

With the exception of some recent results of the Le Fevre and Vine¹⁵ the earlier measurements of the polarizations of carboxylic acids in non-polar solvents made at higher concentrations showed little or no indication that a change in molecular state with further dilution was to be anticipated. Le Fevre and Vine were able to demonstrate an increase in polarizations for mono-, di- and trichloroacetic acids in the lower range of concentrations in which they were able to make measurements. From this they calculated an equilibrium constant for the dissociation of dimer and monomer assuming no other forms were present.

However with reference to their measurements on unsubstituted acids they state "the orientation polarizations of acetic, butyric and hexoic acids only show a slight tendency to increase in most dilute solutions." In this connection a comparison of our results for acetic acid in benzene with theirs is of interest for the present discussion of the precision of our results. This is shown in the lower half of FIGURE 14 in which their three most dilute points (the only ones which overlap ours) have been recalculated to our units and plotted together with our acetic acid results. It will be noted here

¹⁴ Tiganik, L. *Zeit. Physikal. Chem. B* 13: 425. 1931.

¹⁵ Le Fevre, R. J. W., & Vine, H. *Jour. Chem. Soc.* 1795. 1938.

again that the scattering of our points from the best curve through them in the concentration range 2×10^{-4} to 3×10^{-3} is little if any greater than the scattering of their points in the concentration range above 4×10^{-3} .

Consideration of the present results should include the possibility

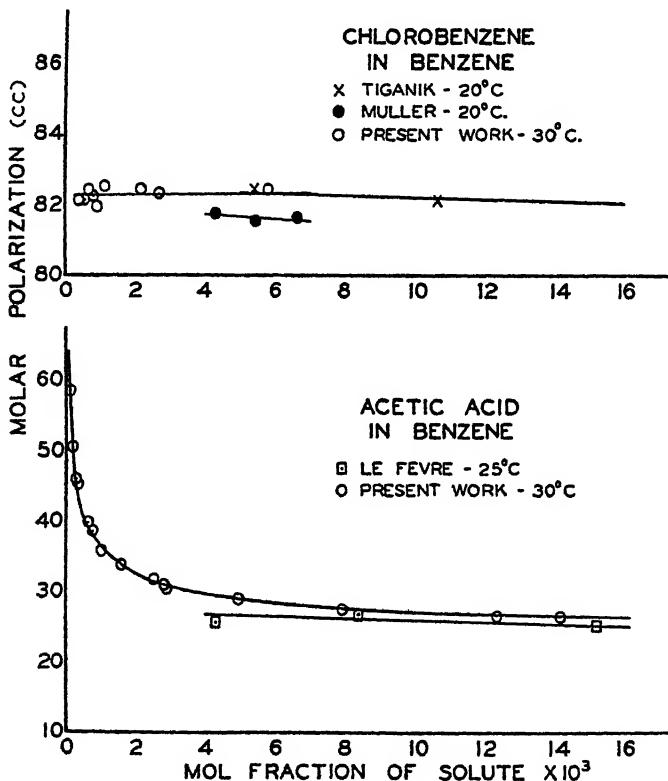


FIGURE 14. Comparison of measurements of Chlorobenzene and Acetic Acid with those of other investigators.

that as a particular dilution is reached the polarization curves would, because of systematic errors in apparatus, or method, or through contamination with traces of moisture, begin to show a false upward trend in the case of the carboxylic acids which was absent for the chemically different molecule chlorobenzene. A comparison of the results for acetic acid in benzene and heptane shown in FIGURE 8 indicates that this is unlikely as the rise in polarization occurs at widely different concentrations in the two cases. The upward curva-

ture found may thus be considered real. Its most likely explanation seems to be the increased concentration of monomer molecules of higher moment from dissociation, as dilution increases, of dimer molecules of lower moment. This is in accord with Lassettre's¹⁶ conclusion, "the organic acids polymerize in non-polar solvents, largely into double molecules . . ."

As a basis for a quantitative interpretation it will be assumed that only the dimer and monomer forms are present in the range of concentrations under consideration and that an equilibrium exists between these which can be expressed as follows:

$$K_a = \frac{A_s^2}{A_D}$$

Where A_s and A_D are the activities of the single and double molecules.

If the activities are assumed proportional to C , the number of moles per unit volume, we obtain:

$$K_a = \frac{C_s^2}{C_D}$$

Now if f_2 is the mole fraction of solute computed on the basis of the monomer molecular weight and f_1 that of the solvent on the same basis, d_{12} the density of the solution and g_s the fraction of the above number of moles (computed as singles) present that are actually single molecules and g_D the corresponding fraction of the molecules computed as singles that are present as double molecules, then we have $g_s + g_D = 1$ and,

$$C_s = g_s \cdot f_2 \cdot \frac{d_{12}}{f_1 M_1 + f_2 M_2} \quad \text{and} \quad C_D = \frac{g_D f_2}{2} \cdot \frac{d_{12}}{f_1 M_1 + f_2 M_2}$$

since $\frac{f_1 M_1 + f_2 M_2}{d_{12}}$ is the molar volume of the solution. We may then obtain,

$$K_a = \frac{2g_s^2}{g_D} \cdot \frac{f_2 \cdot d_{12}}{f_1 M_1 + f_2 M_2} \quad (6)$$

It may be shown analytically that if calculation of P_2 is made according to equation (1) of the introduction in the form

$$P_{12} = f_1 P_1 + f_2 P_2$$

where P_{12} is the polarization per mole of mixture, P_1 is that of the solvent and P_2 is the polarization per mole of solute calculated all as

¹⁶ Lassettre, E. N. Chem. Rev. 20: 301. 1937.

single molecules and f_1 and f_2 are the corresponding mole fractions, the resulting value of P_2 is the average polarization per mole of solute, computing all solute molecules as single. Then the value of P_2 is made up of weighted contributions from the single and double molecules present, as follows:

$$P_2 = g_s P_s + \frac{g_D}{2} \cdot P_D$$

where P_s is the polarization per mole of single molecules and P_D that per mole of double molecules.

Since $g_s + g_D = 1$

$$g_s = \frac{P_2 - \frac{P_D}{2}}{P_s - \frac{P_D}{2}} \text{ and } g_D = \frac{P_s - P_2}{P_s - \frac{P_D}{2}}$$

By substitution in (6) we obtain

$$K_c = \frac{\left(P_s - \frac{P_D}{2} \right)^2 \cdot 2 f_2}{\left(P_s - \frac{P_D}{2} \right) (P_s - P_2)} \cdot \frac{d_{12}}{(f_1 M_1 + f_2 M_2)} \quad (7)$$

The last term of this equation is the reciprocal of the molar volume of the solution. In the dilute solutions here under consideration this is to a close approximation equal to the molar volume of the solvent and may be regarded as constant. With this approximation equation (6) becomes

$$K = \frac{\left(P_2 - \frac{P_D}{2} \right)^2 \cdot 2 f_2}{(P_s - P_2) \left(P_s - \frac{P_D}{2} \right)} = \frac{g_s^2 \cdot 2 f_2}{1 - g_s} \quad (8)$$

It is advantageous to define a new constant K_0 by

$$K_0 = \frac{\left(P_2 - \frac{P_D}{2} \right)^2 \cdot 2 f_2}{(P_s - P_2)} \text{ and hence to write from (8)}$$

$$K_0 = \frac{K}{2} \left(P_s - \frac{P_D}{2} \right) = \frac{\left(P_2 - \frac{P_D}{2} \right)^2}{(P_s - P_2)} \cdot f_2$$

Expansion of this leads to

$$P_s K_0 = P_2^2 \cdot f_2 - P_2 f_2 \cdot P_D + f_2 \left(\frac{P_D}{2} \right)^2 + P_2 K_0 \quad (9)$$

In this equation, there are three unknown constants, P_s , P_D and K_0 and we may solve for these if we use three points from the experimental data for P_2 vs. f_2 . The solution was actually carried out by using a number of points taken from the best curve drawn through the experimental points. Each P_2 vs. f_2 value was set up in an equation of the expanded form and these equations summed up in groups of two, three or four, to give three final equations which were subsequently solved for P_D and K_0 and then for P_s . The method was applied to the data for each acid. The values found for P_s , P_D and

$$\frac{2K_0}{P_s - \frac{P_D}{2}}$$
 are given in TABLE 12.

TABLE 12
CONSTANTS OF ACIDS

Acid	$K \times 10^4$	P_s	P_D	$P_{S(E)}$	$P_S - P_{S(E)}$	$P_D - 2P_{S(E)}$	μ_S	μ_D
Formic.....	9.4	72.6	39.4	8.3	64.3	22.8	1.77	1.06
Formic*.....	<0.08	(70)	32.1	8.3	(62)	15.5	(1.75)	0.87
Acetic.....	3.2	67.3	43.0	13.0	54.3	17.0	1.63	0.91
Acetic*.....	<0.04	(60)	43.8	13.0	(47)	17.8	(1.5)	0.93
Propionic.....	2.3	75.7	51.1	17.6	58.1	15.9	1.68	0.88
Butyric.....	0.72	110	62.2	22.2	78	17.8	1.9	0.93
Trimethylacetic..	0.60	100	73.2	26.8	73	19.6	1.9	0.98
Benzoic.....	4.4	87.4	91.0	32.4	55.0	26.2	1.64	1.13
Monochloroacetic	9.2	124.8	114.8	17.7	107.1	78.8	2.29	1.96

* Heptane measurements; all others in benzene.

The close relation of the computed values of the constants of equation (8) to the experimental results is shown clearly in TABLE 13 where we have tabulated the values of the polarization P_2 for acetic and benzoic acid, computed by the use of the constants K , P_s and P_D under the column P_K . In the last column under $P_{exp.}$ are given the values taken from the best drawn curve at the same mole fractions as used to calculate P_K . Comparison of the differences between the values of P_K and $P_{exp.}$ and the deviations of the observed points from the best experimental curves of FIGURES 8 and 12 will show that these differences are, on the whole, less than the average deviations of the experimental points themselves. It is thus proper to say that the equation adequately represents the relation between P_2 and f_2 .

This of course does not exclude the possibility that some other equation based on a different assumption as to the molecular state of the acids in these solutions might not also represent the observed points adequately. However, in view of the related evidence as to molecular state, particularly from freezing point data, it seems justifiable to state that the monomer-dimer equilibrium we have assumed is substantially correct.

Before proceeding to a consideration of the significance of the

TABLE 13

ACETIC ACID IN BENZENE		
$f_2 \times 10^3$	P_K	$P_{Exp.}$
0.06	57.1	58.5
0.10	53.5	54.5
0.20	48.2	46.4
0.40	42.8	41.1
0.80	38.0	37.4
1.60	34.0	33.8
3.20	30.7	30.4
6.40	28.3	28.1
12.80	26.4	26.7
BENZOIC ACID IN BENZENE		
0.12	75.5	77.5
0.20	72.1	71.9
0.60	64.3	63.4
1.00	61.1	60.7
2.00	57.2	57.4
4.00	54.2	54.1
7.00	52.2	52.1
9.00	51.5	51.5
10.00	51.2	51.3

monomer and dimer polarization values, P_s and P_D for all the acids listed in TABLE 12 it is desirable to emphasize the limitations of these values. In evaluating the constants of equation (8) it is found that the value of P_s obtained is rather sensitive to the choice made for K and vice versa. On the other hand the value found for P_D is relatively insensitive to the choice of P_s and K . This is apparent from the form of the curves themselves. The values of P_D therefore have greater reliability than the values of P_s .

In TABLE 14 a comparison of the present results with those of previous workers is made. On comparison of μ_s values for acetic and propionic acid with those of Zahn, measured in the gas, it is seen

that there is good agreement if allowance is made for the solvent effect demonstrated by Müller's work.¹² A 6 to 10% apparent increase in moment might be expected in going from solvents like benzene and heptane to the gas phase.

Consideration of our results for the monomer moment of formic acid together with the gas values of Zahn, and those of Coop, Davidson and Sutton¹⁷ raises doubt as to the true value of the moment of formic acid. While the high value found by Wilson and Wenzke¹⁸ in dioxane is probably explicable on the basis of some type of co-ordination with the solvent the magnitude of the difference is sur-

TABLE 14
COMPARISON OF DATA

Acid	Zahn ²⁰ (gas)	Sutton ^{*17} (gas)	Briegleb ²²	Wenzke ¹⁸	Wolf ²¹	Present Work Benzene	Work Heptane
Formic							
μ_s	1.5	1.3		2.07		1.77	1.75
μ_D		0.9	1.44		1.77	1.06	0.87
Acetic							
μ_s	1.73			1.74		1.63	1.50
μ_D			1.15		1.13	0.91	0.93
Proxionic							
μ_s	1.74			1.75		1.68	
μ_D			1.02		1.02	0.88	
Benzoic							
μ_s				1.75		1.64	
μ_D		0.79				1.13	

* The values of μ_s and μ_D were calculated from P_S and P_D values given by Coop, Davidson and Sutton by using $\mu = 0.0127\sqrt{(P - P_E)T}$.

prising. In this connection perhaps too little attention has been given to the possibility of difficulties arising from instability of the formic acid. Coolidge¹⁹ points out that the acid stored in glass decomposes slowly even at room temperature. It might be anticipated that decomposition might be greater on metal condenser surfaces especially at elevated temperatures. It is apparent that more work is needed to fix the value of the monomer moment in this case.

The values found for the polarizations of the dimeric forms from

¹⁷ Coop, I. E., Davidson, N. B., & Sutton, L. E. Jour. Chem. Phys. 6: 905. 1938.

¹⁸ Wilson, C. J., & Wenzke, H. H. Jour. Chem. Phys. 2: 546. 1934.

¹⁹ Coolidge, A. S. Jour. Am. Chem. Soc. 50: 2166. 1928.

²⁰ Zahn, C. T. Phys. Rev. 37: 1516. 1931.

²¹ Wolf, K. L. Physikal. Zeit. 31: 227. 1930.

²² Briegleb, G. Zeit. Physikal. Chem. B 10: 205. 1930.

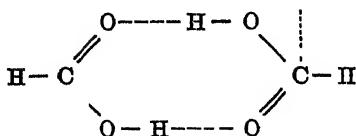
TABLE 12 are for comparative purposes expressed as apparent dipole moments in TABLE 14. Whether the dimer has any real moment will be considered later. Our apparent moments for the double molecules are seen to differ appreciably from those of Briegleb²² and those of Wolf²¹ but are reasonably consistent among themselves. The concentration range for the measurements of these workers was higher than the range of the present work. The work of Smyth and Rogers²³ which covered the concentration range to unit mole fractions of acid, over a range of temperatures, shows that in more concentrated solutions the polarizations increase. The high moments for the dimer found by Briegleb and Wolf are therefore doubtless in part attributable to the presence of higher molecular aggregates than dimers. They were furthermore unable to allow properly for possible contributions to the polarization from the presence of some of the monomer. In the present work this has been done. Our measurements, even at the highest mole fractions still remain in such a low concentration range that significant contributions from higher aggregates appear to be unlikely. It therefore seems justifiable to regard the present values as the most reliable ones for the dimer form of the molecules.

Examination of the polarization values in column 7 of TABLE 12 for the unsubstituted aliphatic acids, with the exception of formic acid in benzene, reveals interesting regularities. If the values for the electron polarizations of the double molecules are subtracted from the corresponding total polarizations the values for the remainder of the polarization are substantially constant and of the order of 15 to 17 cc. Ordinarily this remainder would be classed as orientation plus atom polarization and from this a value for a moment would be computed either by neglecting the atom polarization or by arbitrarily estimating it as some fraction of the sum of the two. The moments so computed would have values of from 0.9 to 1.0 D (column 9 of TABLE 12). However, from our knowledge of the probable structure of the dimer forms of these acids there is reason to suppose that they would have only a very small or zero moment.

The best evidence bearing on the dimer structure is from electron diffraction investigations on formic acid by Pauling and Brockway.²⁴ They conclude that the dimer is an eight membered, hydrogen bonded ring with the following structure:

²² Smyth, C. P., & Rogers, H. E. Jour. Am. Chem. Soc. 52: 1824. 1930.

²³ Pauling, L., & Brockway, L. O. Proc. Nat. Acad. Sci. 20: 339. 1934.



In this structure the eight atoms of the ring are supposed to be in one plane with the angle between the $C = O$ and $C - O$ bonds about 125° . Using the electron diffraction bond distances and angles and Smyth's²⁵ bond moment values and assuming the OH in the position shown a resultant moment for each half of the planar structure can be calculated. When this is done two moments each of value about $1D$ are obtained whose approximate directions are indicated by the two dotted lines directed away from the two carbon atoms in the figure. As these resolved vectors are parallel and oppositely directed the calculated net moment for the dimer is zero.

If the above structure is correct the 15–17 cc. of residual polarization must have some other origin than an appreciable permanent dipole moment in the dimer. In this connection it may be noted that symmetrical distortions of the two halves of the molecule, such as bending in hinge-like fashion about a line passing across the two hydrogen bonds as axis, will not produce a large enough resultant moment to account for the observed polarizations.

The dimer structure is an eight membered ring containing large component bond moments whose net resultants practically cancel and is therefore quite similar to the ring structures of molecules of the type of beryllium acetylacetone, recently measured by Coop and Sutton.²⁶ They have shown that a number of such molecules have total polarization values considerably larger than their electron polarizations and that the total polarization remained constant over a range of temperature. These measurements were made in the gas phase, and were free of solvent complications and thus seemingly indicate the absence of a permanent moment. Coop and Sutton attribute the large polarization values found to atom polarization. The dimer forms postulated for the acids here under investigation have structures which would fit the conditions that they stipulate as necessary for a molecule to show large atom polarization. This therefore seems the most probable explanation at present available of the fairly constant additional polarization value of about 15–17 cc. which we have found for the dimer in the case of acetic, propionic,

²⁵ Smyth, C. P. Jour. Phys. Chem. 41: 209. 1937.

²⁶ Coop, I. E., & Sutton, L. E. Jour. Chem. Soc. 1281. 1938.

butyric and trimethyl acetic acids in benzene solution and for formic acid in heptane. Since all the dimers contain the same ring each acid would probably have about the same amount of atom polarization. The origin of the larger atom polarizations found for formic acid (23 cc.) and benzoic acid (26 cc.) in benzene solution is not at present clear. The polarization of 79 cc. in excess of the electron polarization found for the dimer of monochloroacetic acid is of course largely due to such orientation polarization contributions from the carbon chlorine bond moments as are perpendicular to the long axis of the double molecule. A rough estimate of both the London and dipole interactions between these carbon-chlorine atom pairs and the component atoms and dipoles of the ring indicates that there is probably no free rotation of the $C-Cl$ dipoles about this axis. The favored minimum energy positions would be those perpendicular to the plane of the ring with more molecules present in the form in which the positions of the chlorines are trans to each other. The orientation polarization value computed on this assumption when added to an atom polarization of about 16 cc. for the ring approximates reasonably well to the observed value of 79 cc.

A possible alternative explanation of the additional polarization of 15–17 cc. in these dimer forms is the stepwise breaking of the ring involving rupture of only one of the hydrogen bonds. This would not give any different result from a thermodynamic point of view with respect to the equilibrium constants, nevertheless it would affect the magnitude of the polarization values since within such singly bound dimers there would no longer be complete compensation of the component dipoles and configurations of high moment would probably result as the two halves rotated about an axis along the remaining hydrogen bond. However, unless the fraction of such singly bound dimer molecules, with real orientation polarization, was a constant fraction of the total of the dimer present a fairly constant contribution to the polarization would not be expected. That this fraction would be the same for all of the acids seems improbable in view of their large differences in stability in solution as indicated by the variations in the values of the equilibrium constants which have been found. Furthermore the fraction in the broken form would, if calculated on the basis of a Boltzman factor, be too small to account for the observed residuals.

The present measurements were all made at constant temperature therefore it is not possible to draw quantitative conclusions with regard to the energy change accompanying the dissociation of the dimer

to monomer. The existence of large differences in the extent of dissociation of the different acids is apparent from an inspection of FIGURE 15 where all of the curves are plotted together for comparison. It is clear that the nature of the radical attached to the carboxyl group has an important influence on the dissociation of the dimer ring.

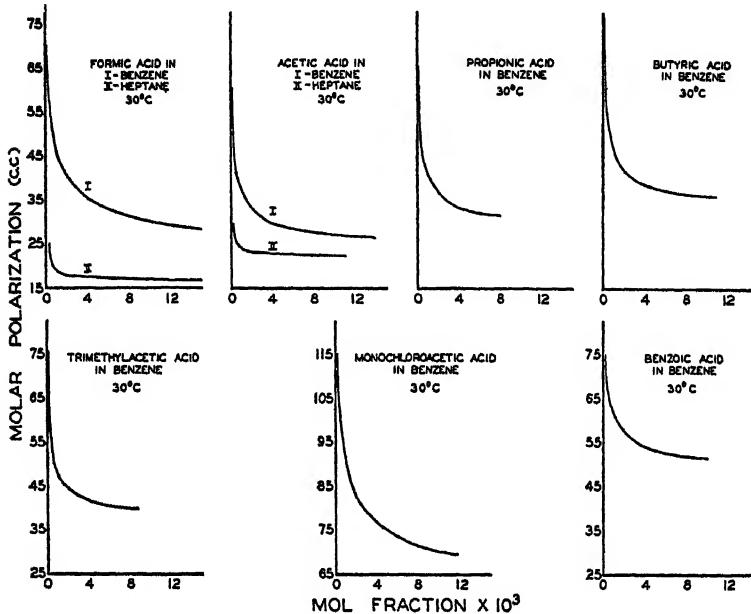


FIGURE 15. Calculated curves for all acids measured.

Likewise the solvent exerts a very marked influence on the extent of dissociation as shown by the much greater dissociation of acetic acid in benzene than in heptane. While tentative explanations of both effects have been formulated their proper interpretation must await the results of further measurements now in progress involving similar polarization determinations at other temperatures and in different solvents.

THE ELECTRICAL PROPERTIES OF POLYVINYL CHLORIDE PLASTICS*

BY RAYMOND M. FUOSS

From the Research Laboratory of the General Electric Co., Schenectady, New York

INTRODUCTION

In recent years, a large number of synthetic high polymers have become available. Their properties are of interest, not only for themselves, but also because a study of them may give some light on the structure of the naturally occurring substances of high molecular weight. The polyvinyl compounds are especially suited for investigation, because Marvel¹ has shown that they are simple chain molecules, with the repeating structure



By choosing polyvinyl chloride, a dipole "handle" is located on every other carbon atom in the chain, thus giving a molecule which will respond to an electric field, and in which the location of the dipoles is known. This is much more information than is available for the high molecular weight compounds of nature, such as the proteins, for example.

Polyvinyl chloride as ordinarily obtained is a white amorphous powder which, under a pressure of about 2000 lbs./sq. in. at 110°, becomes a hard, brittle, transparent solid. The addition of many organic compounds (esters, ethers, ketones, aromatic hydrocarbons) softens or plasticizes the polymer; depending on the relative proportions of the two components, compounds between glass-like substances and very viscous liquids can be obtained, with a rubber-like and a gel stage as intermediates. In this system a very wide range of viscosity is available, in which the properties of electrolytes and dipolar molecules can be studied.

It is the purpose of this paper to present a description of experimental methods of measuring the electrical properties of plastic solids, and to give a preliminary account of the properties of polyvinyl chloride plastics as functions of frequency, temperature and composition.

* Presented in part before the Section on Physics and Chemistry, April 15, 1939 in New York.

¹ Marvel, C. S., & Levesque, C. L. Jour. Am. Chem. Soc. 60: 280. 1938. Marvel, C. S., & Denoon, C. E. Jour. Am. Chem. Soc. 60: 1045. 1938.

EXPERIMENTAL METHODS

Electrical Equipment

In all, four bridges were used. The d.c. conductance was determined on a bridge which permitted 0.1% precision on resistances of 10^6 ohms or less, and progressively lower precision for higher resistances. Bridge voltage was obtained from a set of B-batteries, and could be varied in steps up to 300 volts. Balance was indicated by a galvanometer in the output circuit of a d.c. amplifier across the midpoints of the bridge.²

For samples of high resistance, the dielectric constant and conductance were determined on a Schering bridge, in which the standard was a $100\mu\text{uf}$ quartz insulated three terminal capacitor. For samples of lower resistance ($\tan \delta > 1$), the electrical properties were determined on a resistance bridge, in which the standard was a special 100,000 ohm resistor, non-inductive, wound with manganin, and capable of dissipating 1000 watts without appreciable change in resistance.^{2, 3}

These two a.c. bridges were supplied with voltage by a system of alternators and transformers, which gave up to several kilovolts over the range 15–500 cycles, up to 10 kilovolts over the range 500–2000 cycles, and up to 25 kilovolts at 60 cycles.

Further measurements in the audio frequency range (500–10000 cycles) at low voltage (maximum 10 volts) were made using a parallel bridge, in which the unknown was measured parallel to a 100000 ohm coil in turn parallel to $350\mu\text{uf}$ in a bridge network.⁴ The null instrument for balance on this bridge was a cathode-ray oscilloscope.⁵

Cells and Contacts

A variety of cell types were investigated; the most satisfactory was the one shown in FIGURE 1. It is essential that three terminal cells be used with solids, in order to eliminate edge effects such as fringing and surface conductance as sources of errors. Furthermore, the use of a guard circuit permits calculation of the cell constant from the dimensions of the sample and the test electrode, because at balance, both test electrode and guard ring are at the same potential, and there is therefore normal flux over the area of the test electrode.⁶

²Fuoss, R. M. Jour. Am. Chem. Soc. 60: 451. 1938.

³Fuoss, R. M. Jour. Am. Chem. Soc. 59: 1703. 1937.

⁴Mead, D. J., & Fuoss, R. M. Jour. Am. Chem. Soc. 61: 3589. 1939.

⁵Lamson, H. W. General Radio Experimenter. 13: 5. 1939.

⁶Fuoss, R. M. Trans. Electrochem. Soc. 74: 91. 1938.

The most important difference, from the experimental point of view, between measurements on solids and on liquids is that the latter give adequate contact with the electrodes merely by immersion, while the solids require special precautions. Imperfect contact is to be expected, in general, between solids even with the most carefully machined surfaces, and it was found necessary to prepare the surface

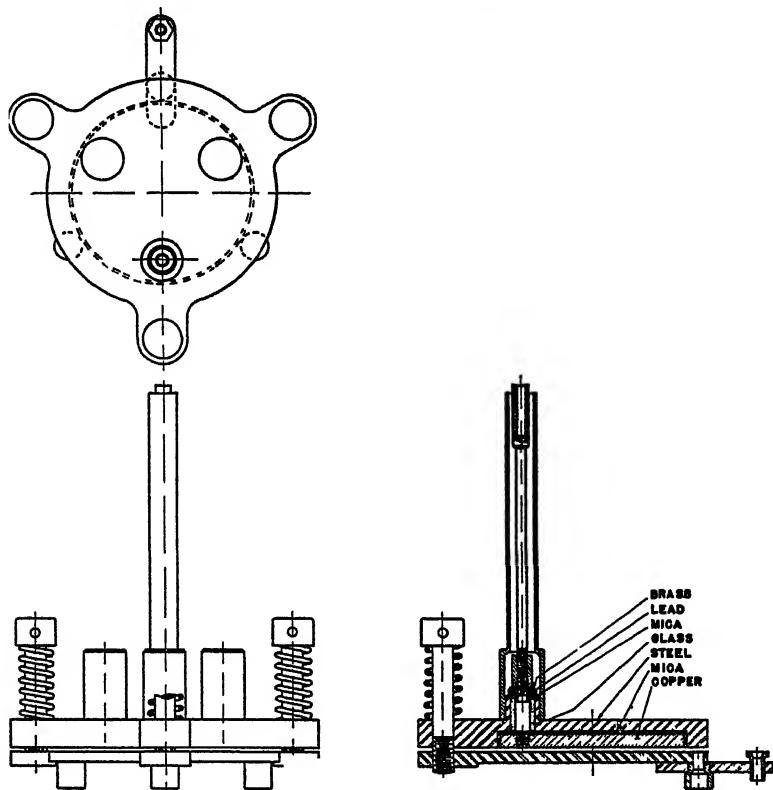


FIGURE 1. Conductance-Capacitance Cell for Solids.

of the sample in some way, in order to eliminate the effects of the series capacity produced by air films. These films were of the order of a few thousandths of a millimeter in average thickness for ground copper electrodes in contact with a fairly soft plastic (40% plasticizer), as determined by measuring the apparent dielectric constant at several thicknesses and frequencies. At low measuring frequencies, a thin air film can produce apparent dielectric constants several times as large as the actual dielectric constant of the unknown.^{5,6} The

criterion for the elimination of surface effects is, of course, independence of volume properties, such as dielectric constant and loss factor, on the thickness of the sample. Painting the surfaces of the sample with diluted aqua-dag, rubbing on powdered graphite, or rubbing on tin foil electrodes (using a minute amount of petrolatum as adhesive in the latter case), gave surface contacts, with which the volume properties were independent of thickness for samples ranging from 1 mm. to 1 cm. in thickness. Depending on the stiffness of the sample, one or another of the above three methods was used.

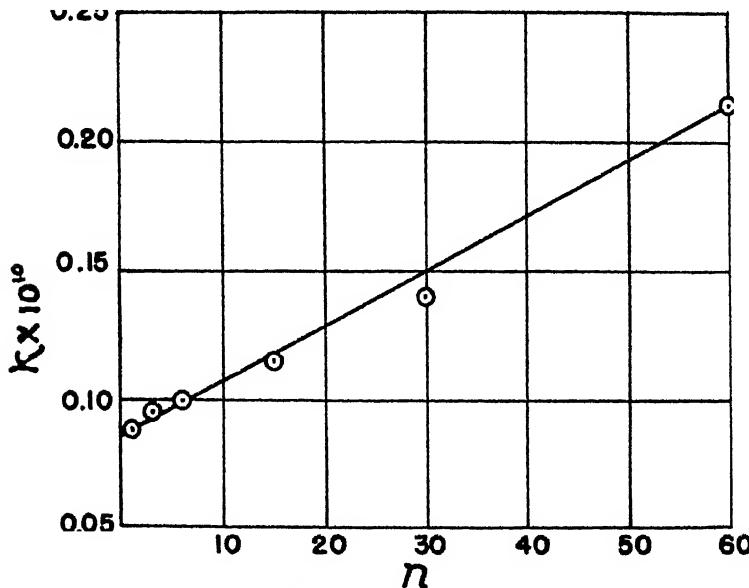


FIGURE 2. Dependence of d.c. Conductance on Milling.

For high frequency work, only the foil method is applicable, because the series resistance of the carbon electrode then becomes a source of error. But up to at least 10000 cycles, this effect is negligible for carbon films on low conductance samples.

For a cell of the type shown in FIGURE 1, the geometrical cell constant C_g is given by

$$C_g = 0.08842 \times 10^{-12} \frac{A}{d} \left(1 + \frac{3}{4} \frac{\Delta r}{r} \right) \quad (1)$$

where A is the area of the test electrode and d is the thickness of the sample. The factor in parenthesis corrects for the slight fringing in the annular space between guard ring and test electrode: Δr is their

distance apart, and r is the radius of the test electrode. This factor was determined by approximate calculation and checked by measuring samples in cells with different values of Δr . The numerical constant in (1) is, of course, the reciprocal of $4\pi \times 9 \times 10^{11}$.

Preparation of Samples

The general method consisted in mixing weighed amounts of the components (plasticiser and polymer) cold as thoroughly as possible, and

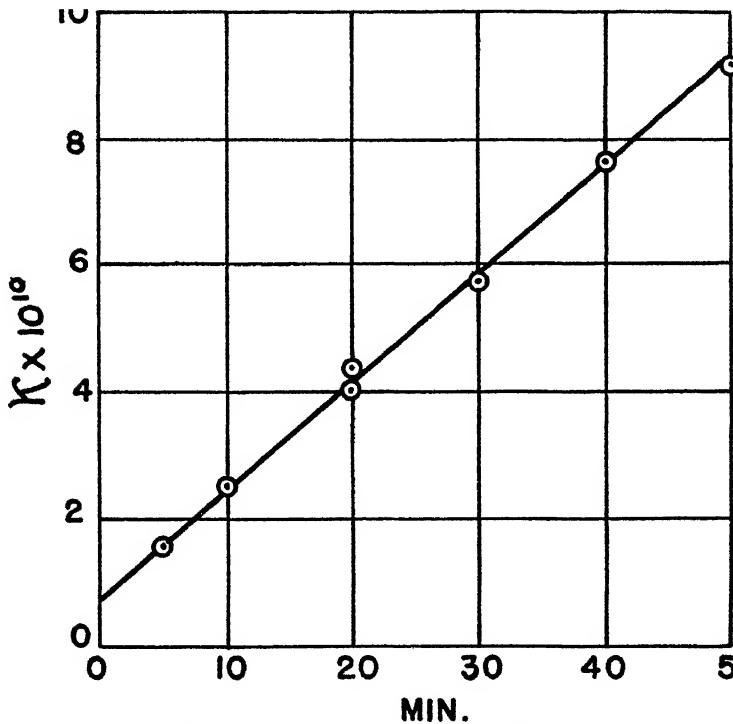


FIGURE 3 Dependence of d.c. Conductance on Pressing.

then hot-mixing them by running the mixture through a small rubber mill at 105°. (The rolls are 6" in diameter and 12" long; one travels at about 22 rpm and the other at 30, so that a shearing mixture occurs between the rolls, where the material is forced to pass. The space between rolls was about 0.5 mm.) The action of the mill produces a thin sheet of plastic from the damp powder fed in. The sheet is put through the mill a number of times, folding and crossing direction

between millings, so that a uniform plastic can be obtained. The sheet from the mill exhibits "elastic memory"; that is, if heated, it will shrink more parallel to the milling direction than across it. Consequently for preparation of the final samples for measurements,

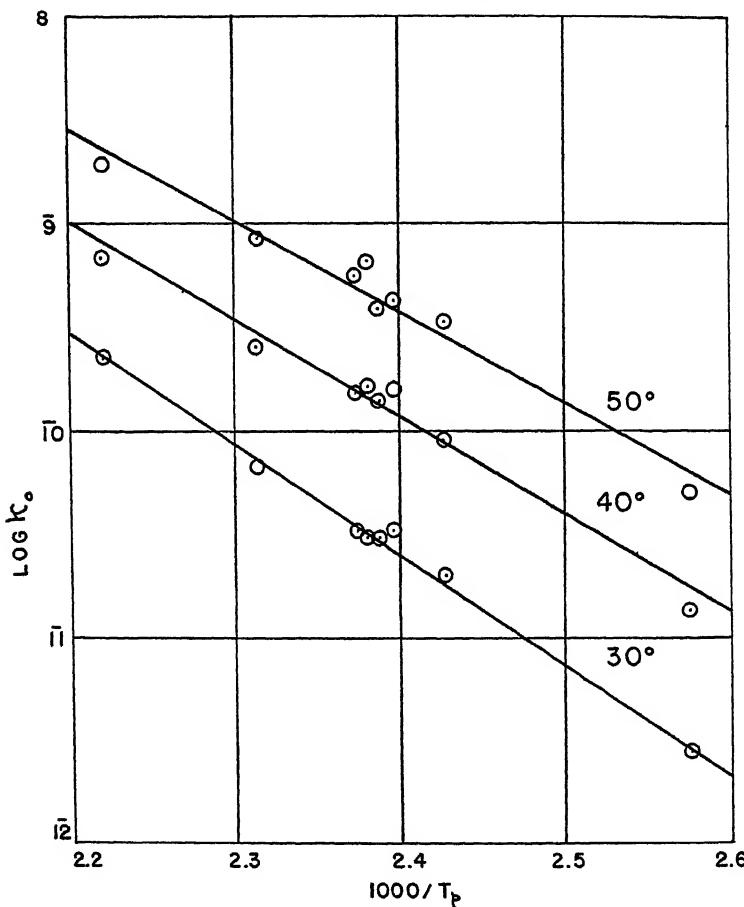


FIGURE 4. Dependence of d.c. Conductance on Press Temperature for three Measuring Temperatures.

15–20 discs were stamped from the mill sheet, stacked with successive mill directions crossed, and pressed at 2000 lbs./sq. in. in a closed hot mold for various recorded times and temperatures. (As will be shown later, the electrical properties of polyvinyl chloride plastics are very sensitive to thermal history, which, therefore, must be carefully controlled and recorded.)

The samples were weighed, coated with surface electrodes and immediately put into the cell, which was then immersed in an oil thermostat where the temperature was held to $\pm 0.01^\circ$. It was found that the measured d.c. resistance was quite dependent on the humidity to which the sample was exposed, independent of the nature of the

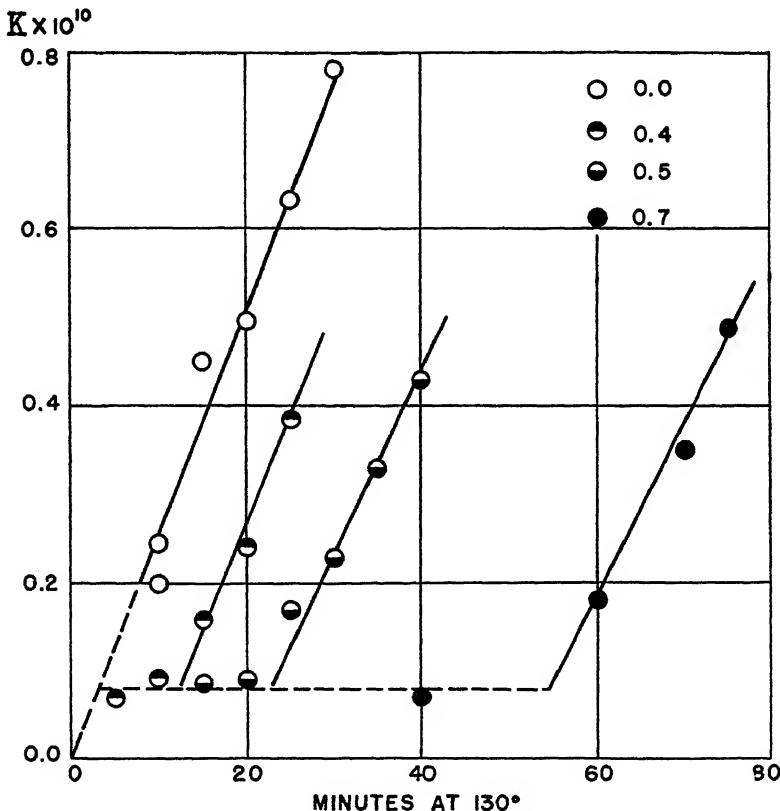


FIGURE 5. Pyrolysis measured by decomposition of lead abietate.

surface electrode. For this reason, the samples were placed in the cell as quickly as possible. The probable explanation is that moisture absorbed from the air increased the conductivity, but more work remains to be done on the effects of water vapor.

The samples also showed another change in electrical properties which seems to be an inherent characteristic of plastics. The dielectric constant decreases with time, approaching asymptotically a value several percent smaller than the initial value. The logarithm of the

difference between the dielectric constant at time t and its asymptotic value is a linear function of \sqrt{t} , with a slope depending on the frequency of the measurement. The nature of the effect suggests a relaxation of internal strains by a diffusion-like mechanism. It should be mentioned that the density remains constant within 0.1% during this change.

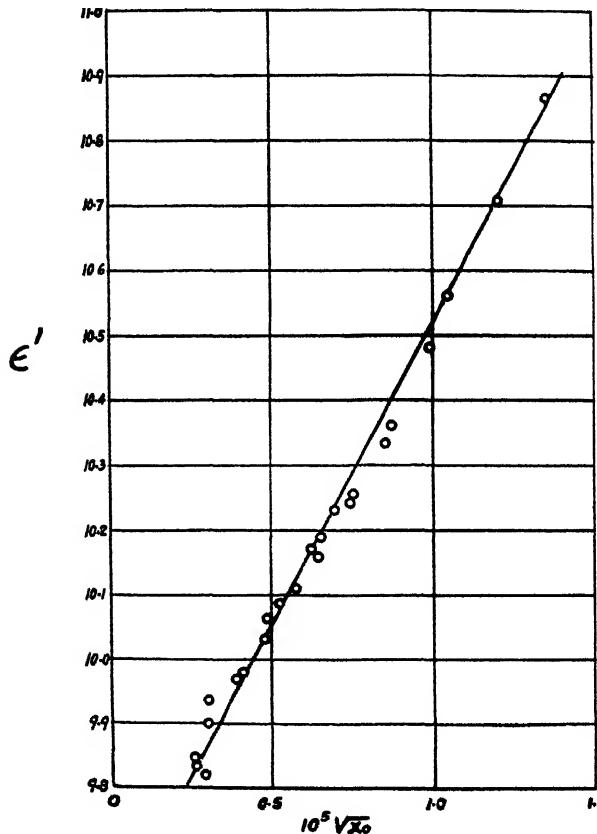


FIGURE 6. Dependence of 60 cycle Dielectric Constant on d.c. Conductance.

RESULTS

Pyrolysis

It was observed, early in the course of this work, that the conductance increased with increasing severity of thermal treatment to which the sample was exposed; a systematic investigation was there-

fore made. It was found that the effect varied from plasticizer to plasticizer; results for only tricresyl phosphate will be presented here.

The conductance increased linearly with the time of heating at a given fixed temperature, as is shown in FIGURES 2 and 3. The first shows the increase in conductance (measured at 40°) due to milling at 105° for a series of samples containing 40% tricresyl phosphate and 60% polyvinyl chloride, which were pressed for 5 minutes at

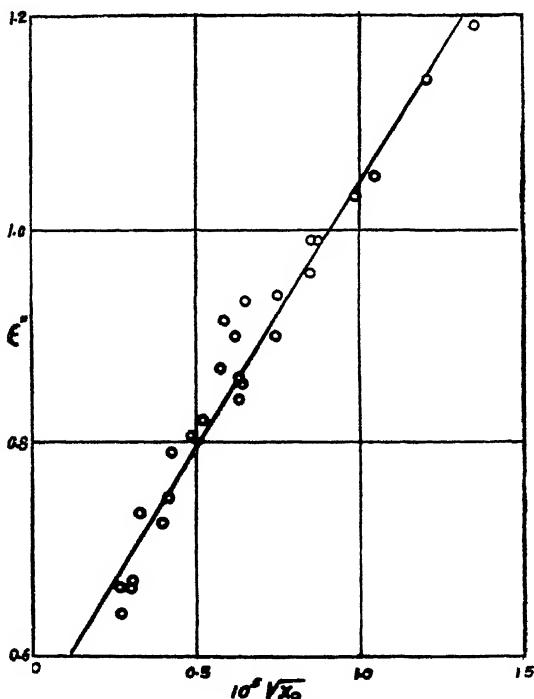


FIGURE 7. Dependence of 60 cycle Loss Factor on d.c. Conductance.

130°. The second shows the increase in conductance due to pressing for various times at 150° samples which were milled 15 times at 105°. In both cases, straight lines are obtained when conductance is plotted against duration of thermal treatment. The intercept at zero milling time in FIGURE 2 corresponds to the conductance produced by pressing a (hypothetical) unmilled sample, and the intercept of FIGURE 3 corresponds to the conductance produced by the milling alone; plus, in both cases, conductance due to original conducting impurities, corresponding to the familiar solvent conductance of ordinary conductimetric work.

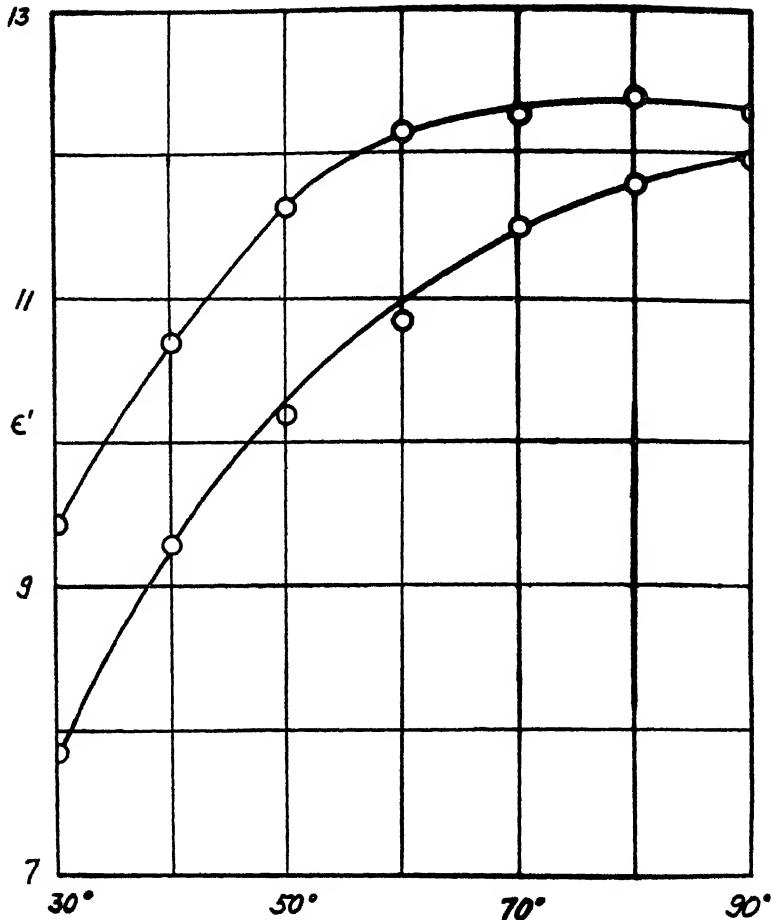


FIGURE 8. Dielectric Constant as a Function of Temperature, upper curve, 60 cycles; lower curve, 500 cycles.

If samples are pressed at different temperatures for the same time, and correction is made for milling, then a straight line is obtained when the logarithm of the d.c. conductance is plotted against the reciprocal of the absolute press temperature, as is shown in FIGURE 4. In the figure, conductance data for three different measuring temperatures are given; it will be seen that the slope varies slightly with measuring temperature. This is because the temperature coefficient of conductance is itself dependent on its electrolyte concentration as measured by its conductance at a given temperature. The conduct-

ance of a given sample is approximately exponential; a plot of the logarithm of the d.c. conductance against reciprocal of absolute measuring temperature is slightly concave down, and the average slope decreases with increasing initial electrolyte concentration.

The effects of thermal history on polyvinyl chloride-tricresyl phosphate plastics can thus be summarized approximately by the following equations:

$$\log \kappa_0 = A - B/T_m \quad (2)$$

$$= C \int_0^t \exp [-D/T(t)] dt \quad (3)$$

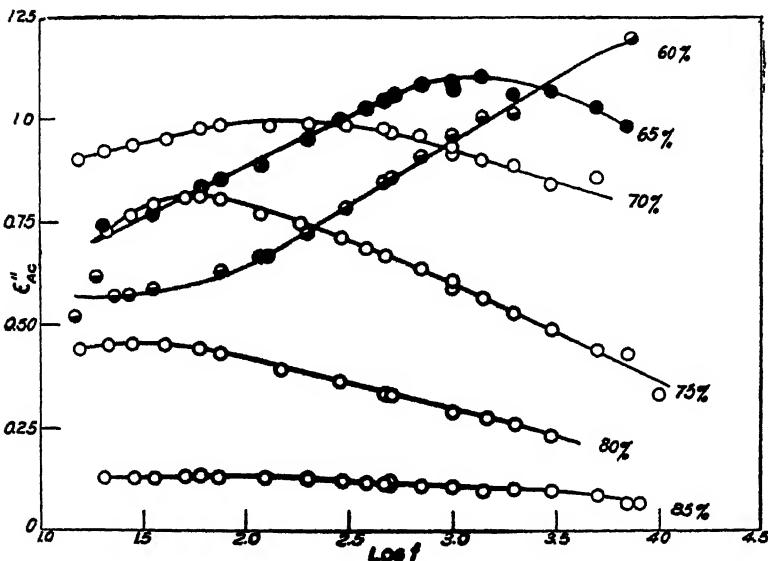


FIGURE 9. Dependence of Loss Factor on Frequency and Composition for polyvinyl-chloride-tricresyl phosphate plastics.

where A depends primarily on the composition of the sample (relative amounts of plasticizer and polymer), B depends on the electrolyte content and composition, C depends on composition and measuring temperature T_m , D depends on measuring temperature and composition, and $T(t)$ represents the temperature at time t during the history of the sample.

It was logical to expect that the conductance was due to hydrogen chloride, liberated from the polyvinyl chloride by pyrolysis, and behaving like a weak electrolyte in the plastic. In order to test this assumption, some samples were made up in which lead abietate in

various amounts were added to the tricresyl phosphate. After pressing for various times, the conductance at 40° was determined. The results of a typical run are shown in FIGURE 5. It will be seen that the curve consists of two parts: an initial horizontal segment, in which the conductance is low and independent of heating time, and a segment in which the conductance increases linearly with heating

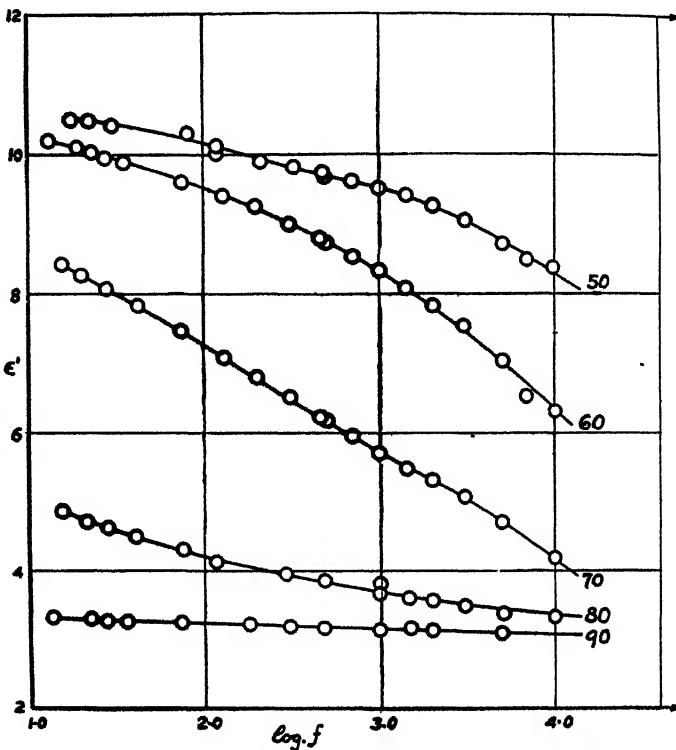


FIGURE 10. Dependence of Dielectric Constant on Frequency and Composition for Polyvinylchloride-Tricresylphosphate Plastics.

time, at a rate equal to that of the blank. The intersection of these two portions of the curve determines the time at which the lead abietate was all converted to chloride; up to this time, no free hydrogen chloride was present.

The a.c. properties depend on the amount of electrolyte present. The d.c. conductance κ_0 is a figure which summarizes the thermal history of a given sample, and a simple empirical correlation has been found in turn between the a.c. properties and κ_0 . Both the

dielectric constant ϵ' and the a.c. loss factor ϵ'' increase linearly with the square root of d.c. conductance, as is shown in FIGURES 6 and 7 for a series of samples containing 40% tricresyl phosphate and 60% polyvinyl chloride. These samples were heated at various times and temperatures in order to cover a wide range of the variables and measured at 40°. The slope of the $\epsilon' - \sqrt{\kappa_0}$ and $\epsilon'' - \sqrt{\kappa_0}$ plots decreases with increasing frequency; the data of FIGURES 6 and 7 were taken at 60 cycles.

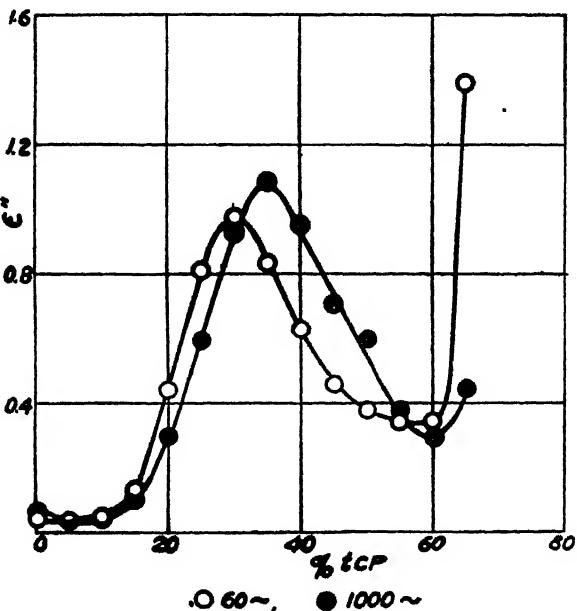


FIGURE 11. Dependence of Loss Factor on Composition at 60 and 1000 Cycles.

It is thus clear that the power absorption in the plastic involves two mechanisms. We can define a total loss factor ϵ''_t as follows:

$$\epsilon''_t = \kappa / 0.08842 \times 10^{-12} \omega \quad (4)$$

where κ is the a.c. conductance and ω is 2π times the frequency. But

$$\kappa = \kappa_0 + 0.08842 \times 10^{-12} \omega \epsilon'' \quad (5)$$

that is, the in-phase component contains the electrolytic (d.c.) conductance and the pure a.c. response, defined by ϵ'' . The latter in turn is made up of two parts, a contribution due to the electrolyte in the plastic and a part due to the plastic itself, corresponding to the intercept at $\kappa_0 = 0$ in FIGURE 7.

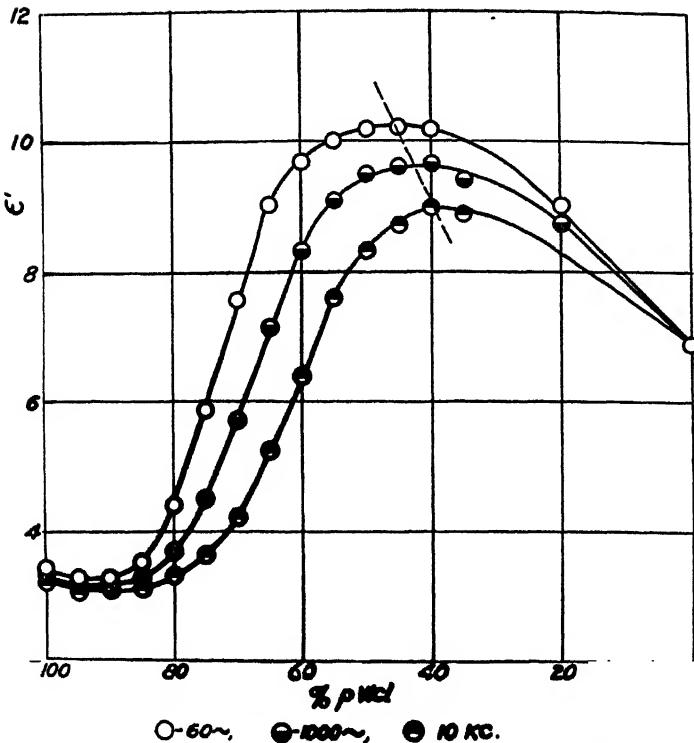


FIGURE 12. Dependence of Dielectric Constant on Composition at 60, 1000 and 10,000 cycles.

DEPENDENCE OF PROPERTIES ON TEMPERATURE, FREQUENCY AND COMPOSITION

With increasing temperature, the d.c. conductance increases exponentially with temperature. The dielectric constant also increases with increasing temperature; FIGURE 8 is a typical $\epsilon' - T$ curve, for a 40-60 plastic measured at 60 and at 500 cycles. In general, the higher the plasticizer content, the less the dielectric constant changes with temperature in the range 30-100°, as might be expected. The a. c. loss factor, ϵ'' , decreases slowly with increasing temperature in the 40-60 plastic at 60 cycles, in marked contrast to the rapid change characteristic of ordinary polar liquids.

In TABLES 1 and 2 are given data which summarize the dependence of dielectric constant and loss factor on frequency and composition at 40° for the system tricresyl phosphate-polyvinyl chloride. Pure polyvinyl chloride is hard and brittle; the dielectric constant is low, as

TABLE 1

DIELECTRIC CONSTANTS OF POLYVINYL CHLORIDE-TRICRESYL PHOSPHATE PLASTICS AT 40°

%PVC _I	<i>Frequency</i>								
	20~	60~	120~	250~	500~	1000~	2000~	5000~	
100	3.454	3.435	3.419	3.397	3.370	3.338	3.306	3.263	3.231
95	3.300	3.268	3.257	3.239	3.223	3.204	3.182	3.154	3.132
90	3.320	3.275	3.250	3.225	3.195	3.170	3.145	3.115	3.085
85	3.640	3.510	3.440	3.375	3.320	3.270	3.220	3.165	3.120
80	4.760	4.395	4.190	4.010	3.840	3.695	3.580	3.425	3.310
75	6.50	5.87	5.48	5.10	4.79	4.50	4.23	3.88	3.64
70	8.29	7.56	7.10	6.62	6.17	5.72	5.27	4.67	4.21
65	9.47	9.02	8.64	8.14	7.65	7.14	6.62	5.82	5.24
60	10.00	9.69	9.42	9.08	8.62	8.32	7.78	6.97	6.37
55	10.33	10.00	9.80	9.59	9.37	9.09	8.75	8.08	7.57
50	10.46	10.19	10.02	9.85	9.68	9.50	9.25	8.76	8.30
45	10.45	10.21	10.06	9.90	9.75	9.60	9.40	9.06	8.70
40	10.38	10.18	10.05	9.92	9.80	9.65	9.48	9.22	8.97
35	(12.25)	(11.00)	(10.38)	(10.88)	9.55	9.40	9.28	9.08	8.86
20	—	8.99	—	—	8.90	8.72	—	—	—
0	—	6.92	—	—	6.89	6.89	—	—	—

TABLE 2

LOSS FACTORS OF POLYVINYL CHLORIDE-TRICRESYL PHOSPHATE PLASTICS AT 40°

%PVC _I	<i>Frequency</i>							
	20~	60~	120~	250~	500~	1000~	2000~	5000~
100	0.031	0.040	0.046	0.053	0.060	0.066	—	—
95	0.033	0.035	0.035	0.037	0.038	0.040	0.042	—
90	0.060	0.058	0.056	0.054	0.052	0.049	0.045	—
85	0.133	0.131	0.127	0.124	0.116	0.107	0.097	0.079
80	0.451	0.441	0.405	0.368	0.330	0.294	0.256	0.208
75	0.752	0.810	0.778	0.726	0.664	0.598	0.530	0.440
70	0.920	0.978	0.994	0.992	0.972	0.930	0.880	(0.820)
65	0.720	0.836	0.904	0.976	1.056	1.090	1.084	1.080
60	0.550	0.630	0.686	0.770	0.860	0.952	1.040	1.160
55	0.45	0.46	0.49	0.55	0.63	0.71	0.82	0.98
50	0.37	0.38	0.40	0.45	0.51	0.60	0.70	0.85
45	0.3	0.345	0.3	0.30	0.34	0.38	0.44	0.55
40	—	0.348	—	0.27	0.29	0.30	0.32	—
35	—	1.39	1.05	0.80	0.60	0.45	0.35	—

is the loss factor, and both change only slowly with frequency. The addition of plasticizer at first decreases the dielectric constant and loss factor; at about 10% tricresyl phosphate, a minimum appears, followed by a very rapid rise in ϵ' and ϵ'' . The steepest portions of the ϵ' curves come in the range 60–80% polymer, where the macroscopic mechanical properties of the plastic are changing most rapidly, and correspond to the rubber-like stage. Simultaneously, the loss factor goes through a sharp maximum, which is not very sensitive to frequency in the present range of frequency. There is also evidence of a second maximum in the gel stage, beyond 60% plasticizer.⁷

The change of ϵ' and ϵ'' with frequency superficially resembles that of ordinary polar systems, in that the loss factor goes through a maximum at the frequency at which the $\epsilon' - f$ curve has an inflection point. There is, however, a striking difference between the plastic and the usual polar liquid: The maximum in the $\epsilon' - f$ curve and the inflection region of the $\epsilon' - f$ curve are extremely broad. For example, in the sample containing 70% polyvinyl chloride, only the tip of the maximum appears in a range covering three decades of frequency, and the inflection region is fully three decades wide, because the $\epsilon' - \log f$ plot is practically linear. At higher polymer content, the $\epsilon' - \log f$ curve is concave up, indicating a center of absorption at frequencies below 15 cycles, the present lower limit of measurement, while samples containing more than 30% plasticizer have centers of absorption at frequencies beyond 10 kilocycles.

There is experimental evidence that the loss factor does not decrease even at very low frequencies. If the a.c. conductance is plotted against frequency in the low frequency range, a practically linear curve is obtained which extrapolates linearly to the d.c. conductance. In any relaxation mechanism, the loss factor must reduce to zero at zero frequency; in the plastics, finite absorption persists to extremely low frequencies. If a relaxation mechanism is involved, a very wide distribution of relaxation times⁸ would be needed to account for the width of the dispersion, and furthermore, some of the times of relaxation would have to be enormous, in order to account for the response at low frequency. The nature of the energy absorption suggests that at least two mechanisms are involved: an ordinary relaxation (possibly with a distribution of relaxation constants), superimposed on a frequency independent static friction.

⁷Fuoss, R. M. Jour. Am. Chem. Soc. 61: 2329. 1939.

⁸Wagner, K. W. Ann Physik. 40: 817. 1913. Yager, W. A. Physics. 7: 434. 1936

SUMMARY

1. Three terminal cells and special surface contacts are recommended for the measurement of the electrical properties of solids.
2. The properties of polyvinyl chloride plastics, especially the d.c. conductance, depend upon the thermal history of the sample.
3. Energy absorption in plastics changes much more slowly with frequency and temperature than in ordinary polar systems.
4. The electrical properties depend on the relative amounts of polymer and plasticizer in the plastic; the dependence is sharpest in the range of plasticity corresponding to the rubber-like stage, intermediate between the glass-like and gel stages.

THE DIELECTRIC CONSTANTS OF SOME ORGANIC CRYSTALS AND GLASSES

BY WILLIAM O. BAKER AND CHARLES P. SMYTH

From the Department of Chemistry, Princeton University, Princeton, New Jersey

Knowledge of intermolecular action in condensed phases has been advanced in recent years by application of the wave mechanics.^{1, 2} New techniques for the measurement and interpretation of molecular structure, particularly dipole moment determinations^{3, 4} and X-ray and electron diffraction⁵ have extended and revised the classical stereochemistry. It has lately become possible to base upon the known shape and size of a given molecule and the electrical properties of its constituent groups, generalizations as to the aggregate properties of crystal form, liquid fluidity, fusion process, entropy, free energy, and the like. Modern statistical mechanics provides a precise approach to this interrelationship,⁶ but computation difficulties lend encouragement to the use of experimental methods. Among these is investigation of the behavior of permanently polar molecules perturbed by an imposed electric field of known and variable frequency. Such recent dielectric studies on solids have frequently revealed a surprising mobility^{7, 8, 9} in a crystalline array previously thought fixed and ordered. This report attempts to summarize results from a dielectric examination of a series of highly purified organic compounds of simple, known shapes which were progressively varied by the use of stereoisomers or consecutive homologues. The experimental method, data, and detailed conclusions are recorded elsewhere.¹⁰

¹ London, F. *Zeit. Physik.* **63**: 243. 1930. *Zeit. Physikal. Chem.* **B 11**: 222. 1931.

² Slater, J. C., & Kirkwood, J. G. *Phys. Rev.* **37**: 682. 1931.

³ Debye, P. "Polar Molecules." Chemical Catalog Company, New York. 1929.

⁴ Smyth, C. P. "Dielectric Constant and Molecular Structure." Chemical Catalog Company, New York. 1931.

⁵ Pauling, L. "The Nature of the Chemical Bond and the Structure of Molecules and Crystals." Cornell University Press, Ithaca. 1939.

⁶ See Fowler, R. H. "Statistical Mechanics." Second Edition. Cambridge University Press. 1936.

⁷ Pauling, L. *Phys. Rev.* **36**: 430. 1930.

⁸ Smyth, C. P. *Chem. Rev.* **19**: 329. 1936.

⁹ Eucken, A. *Zeit. Elektrochem.* **45**: 128. 1939.

¹⁰ Baker, W. O., & Smyth, C. P. *Jour. Am. Chem. Soc.* **60**: 1229. 1938; *Jour. Am. Chem. Soc.* **61**: 1695. 2063. 2798. 1939.

ROD AND PEAR-SHAPED MOLECULES

Information on long-chain, highly anisotropic esters and cetyl alcohol^{11, 12} indicated rotation in the crystal, in a region bounded by fusion and pronounced thermal transitions. Here the librations were about the long axis, yet their loss with lowered temperature resulted in monotropic changes of density, packing and crystal form. Definite 'pre-melting' phenomena were demonstrated. In the present series, these effects were sought in n-amyl bromide, the most extended structure included, for which a thermal transition had already been indicated.^{13, 14} FIGURE 1 shows the dielectric behavior of n-amyl bromide on cooling and warming. It apparently exhibits a monotropic transition resulting in a low temperature form which melts at -88.7° , about 5.9° above the phase which first appears on freezing. It thus resembles diethyl ether.^{15, 16} Skau and McCullough,¹⁴ in confirming the presence of two forms, an unstable one melting at -94.6° (a value close to -94.5° , our higher freezing point indicated on FIGURE 1), and a stable one melting at -87.9° , quote Simon¹⁷ as reporting a freezing point of -95.25° , a value close to the lower freezing temperature on FIGURE 1. The dielectric measurements, tracing the mechanical behavior of the polar molecules, reveal that despite prevention of much supercooling by anisotropy, crystallization either does not proceed rapidly enough, or does not release sufficient latent heat, completely to return the system to a "true" freezing point. Thus is explained the divergent data of previous investigators, since the freezing point may appear to vary with the bath gradient.

Many polymorphic transitions involving only anisotropic lattices, as the polarizing microscope showed both forms of n-amyl bromide to be,¹⁸ are found to involve rotational freedom about at least one molecular axis.¹² However, FIGURE 1 indicates that after occasional orientation just below the freezing point has ceased, only electronic and atomic polarizations contribute to the dielectric constant. An X-ray analysis of the transformation would seem desirable, as the rearrangement in the lattice seems not to occur by the usual librations, and chain translations as proposed by Schoon¹⁹ for lengthy molecules might obtain.

¹¹ Smyth, C. P., & Baker, W. O. *Jour. Chem. Phys.* **5**: 666. 1937.

¹² Baker, W. O., & Smyth, C. P. *Jour. Am. Chem. Soc.* **60**: 1229. 1938.

¹³ Deesse, E. F. *Jour. Am. Chem. Soc.* **53**: 3873. 1931.

¹⁴ Skau, E. L., & McCullough, R. *Jour. Am. Chem. Soc.* **57**: 2439. 1935.

¹⁵ Huetting, H., & Smyth, C. P. *Jour. Am. Chem. Soc.* **57**: 1523. 1935.

¹⁶ McNeight, S. A., & Smyth, C. P. *Jour. Am. Chem. Soc.* **58**: 1718. 1936.

¹⁷ Simon, I. *Bull. Soc. Chim. Belg.* **38**: 47. 1929.

¹⁸ Baker, W. O., & Smyth, C. P. *Jour. Am. Chem. Soc.* **61**: 1695. 1939.

¹⁹ Schoon, T. *Zeit. Physik. Chem. B* **39**: 385. 1938.

Optical examination showed the transition directly. The initial crystals were strongly birefringent, but the double refraction increased when a given sample was held at a low temperature. When the sample was frozen, cooled briefly about 30° below the melting point, and then allowed to warm up quite rapidly, sharp melting of the unstable form was first observed. However, distributed through the liquid were tabular, strongly anisotropic crystals, which did not themselves fuse until a distinctly higher temperature. The latter were in the so-called stable form. On prolonged exposure to low temperatures, the transformation was complete, and only one melting process was visible microscopically.

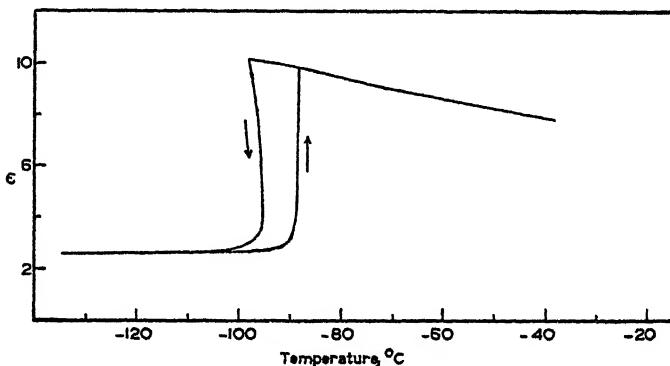


FIGURE 1. Temperature Dependence of the Dielectric Constant of n-amyl bromide.

FIGURE 2 represents the general immobilizing of the dipoles of the more symmetrical i-propyl bromide upon freezing, after slight supercooling and with hindered rotational freedom extending to about -131° . Below this temperature, dispersion has disappeared, and the dielectric constant is nearly the square of the refractive index, as electron polarization alone would require. The polarizing microscope shows pronounced anisotropy on solidification, so rotation is probably not about all axes. The cubic, or, occasionally, the hexagonal, class is characteristic of molecular crystals in which rotational molecular freedom comparable to that in the liquid exists.⁸ The sharp reduction of the dielectric constant at the freezing point implies that rotation around any axis perpendicular to the $C - Br$ line ceases at this temperature, while the gradual decline of the dielectric constant accompanied by anomalous dispersion from -90.8° to about -131° indicates the possibility of a decreasing hindered

libration about this axis or one nearly parallel to it, a behavior undetectable in many structures containing one principal dipole.

Although beginning at the same temperature as that at which it ends on cooling, the dispersion is more pronounced on warming, and a cooperative loosening, like rotational premelting, occurs,^{12, 20, 21} whose effect will be emphasized subsequently. Also, dielectric loss data,¹⁸ confirm the idea of restrained rotation with no transition, since ϵ'' , related to the loss factor, $\tan \delta = \frac{\epsilon''}{\epsilon'}$, where ϵ' is the dielec-

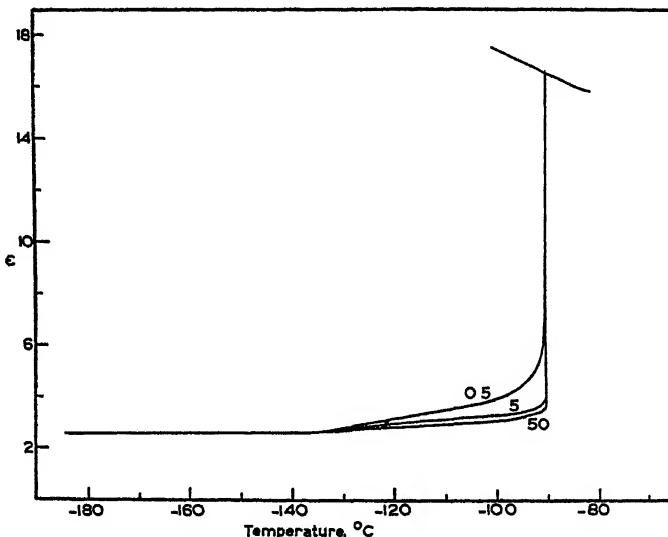


FIGURE 2. Temperature Dependence of the Dielectric Constant of i-propyl bromide.

tric constant, shows no maximum but a regular decrease after solidification.

The i-propyl bromide molecule is sufficiently symmetrical to pack with an economy of free volume in the liquid, but its anisotropic lattice suggests that the liquid arrangement is not of cubic packed spheres. Hence, some supercooling is expected (see later discussion), but soon, on cooling, the thermodynamic potential, probably with the aid of short relaxation times, is able to effect crystallization with free energy diminution, unlike the cases of the bulkier i-butyl and i-amyl compounds, whose vitrification is next considered. Further, standard

¹² Ubbelohde, A. R. Trans. Faraday Soc. 34: 282, 292. 1938.

²⁰ Mueller, H. Proc. Roy. Soc. London. A 158: 403. 1937.

theories that supercooling proceeds in the virtually complete absence of nuclei are supported by the exact continuity of the supercooled ϵ curve with that for the liquid, in FIGURE 2. Minute crystallites would cause a sharp change in the dielectric constant.²²

The vitreous and crystalline behavior of molecules intermediate in symmetry between n-amyl bromide and the spherical t-butyl halides is now to be considered. Uncertainties persist about the thermal and optical properties of glasses, about their mechanism of viscous flow, and the reality of definite vitrification and devitrification temperatures.²³ Commercial glasses, widely used as dielectrics have been explored as to dielectric constant and loss, and breakdown potential, usually over short temperature ranges.²⁴ These, like ordinary organic glasses such as sugars and glycerol, have strong intermolecular bonding, which complicates dielectric interpretations of structure. However, i-butyl and i-amyl bromides, whose hydrogen bond type association is virtually negligible, invariably form clear glasses on cooling.

Previous dielectric measurements on organic glasses are listed in the Landolt-Börnstein "Tabellen."²⁵ The decreasing temperature portions of the curves in FIGURES 3 and 5, on which the appropriate applied frequencies are marked in kilocycles, resemble those of Thomas,²⁶ for liquid and glassy glucose. Similar, more detailed studies on glucose,²⁷ glycerol and several monohydric alcohols²⁸ by Kobeko and co-workers produced the same sort of marked dispersion and decrease of dielectric constant over a short temperature range, whose position was established by the frequency of the field. Loss factors of commercial glasses have been measured at high frequencies by Hackel,²⁹ who used the immersion method of finding a polar liquid mixture with the same dielectric constant as the glass.

Basic postulates of the present report agree with Hägg's³⁰ conclusion from X-ray data that organic glasses contain large groups of molecules organized by more or less associative forces. Glycerol glass has been found truly amorphous³¹ while other X-ray patterns

²² Errera, J. Physik. Zeit. der Sovietunion. 3: 443. 1933.

²³ Richards, W. T. Jour. Chem. Phys. 4: 449. 1936.

²⁴ Morey, G. "The Properties of Glass." Reinhold Publishing Corp., New York. 1938. Chap. XX.

²⁵ See also Tamman, G. "Der Glaszustand." Verlag L. Voss. Leipzig. 1933.

²⁶ Thomas, S. B. Jour. Phys. Chem. 35: 2103. 1931.

²⁷ Kobeko, P. P. et al. Physik. Zeit. Sovietunion. 4: 83. 680. 1933.

²⁸ Kobeko, P. P. et al. Jour. Tech. Phys. U. S. S. R. 8: 715. 1938.

²⁹ Hackel, W. Physikal. Zeit. 37: 160. 1936; Ann. Physik. 29: 63. 1937.

³⁰ Hägg, G. Jour. Chem. Phys. 3: 42. 1935.

³¹ Lark-Horovitz, K., & Miller, E. P. Phys. Rev. 47: 813. 1935.

have been accepted as verification of at least macroscopic non-crystallinity.³²

The graphs of FIGURES 3 and 5 obtained on cooling follow the Debye temperature dependence for a polar liquid down to the regions — 150 to — 165° for i-butyl bromide and — 135 to — 155° for i-amyl bromide. In these ranges occurs a gradual frequency-dependent drop of the dielectric constant to approximately the refraction value. This apparent rotational freezing differs from ordinary rotational immobilization on solidification in at least two ways. It is not

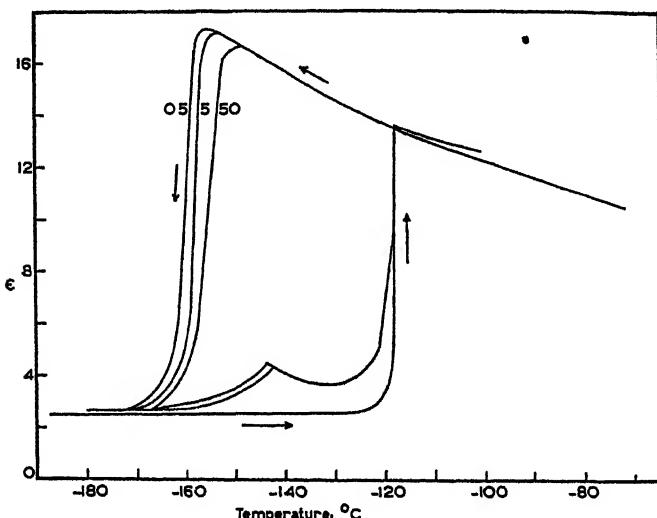


FIGURE 3. Variation with Temperature and Frequency of the Dielectric Constant of i-butyl bromide.

initiated sharply and shows no latent heat under the conditions of measurement. The gradual loss of librational freedom suggests that the individual dipoles are orienting in a viscous medium for, at a given temperature, it is seen that the molecules may be able to follow quite completely a field of frequency 500, and show little or no response in a field of 50,000 cycles.

Mechanically, it was found that i-butyl bromide became stiff at about — 120°, and i-amyl bromide at about — 115°. Seemingly, with decreasing temperature and increasing density, enhanced molecular field interaction, especially from the short-range forces,¹ removes oscillational freedom even in the absence of a phase change.

³² Warren, B. E. Jour. Applied Phys. 8: 645. 1937.

An artificially increased interaction, such as the imposition of high pressures, should then produce the same effect as lowering the temperature. This is affirmed by the experiments of Danforth,³³ who introduced pronounced dispersion at given frequencies in some ten compressible liquids like glycerol and isobutyl alcohol by measuring the dielectric constant under high pressures.

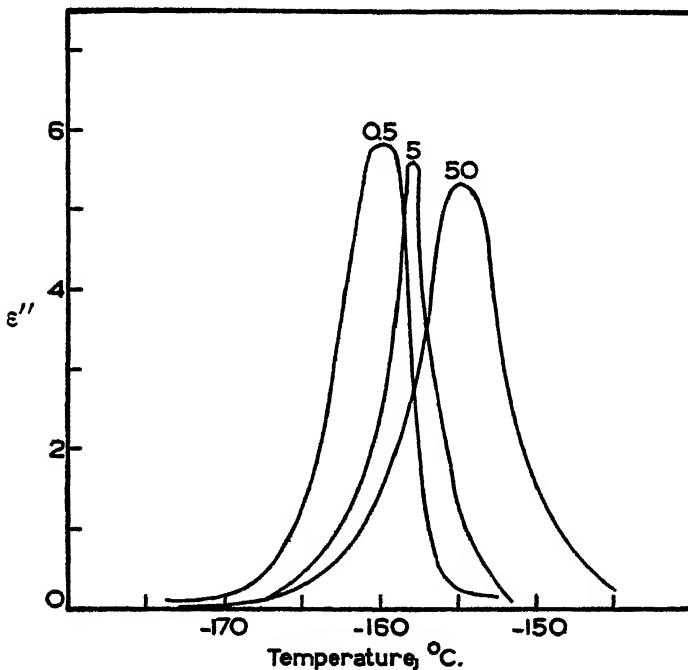


FIGURE 4. Variation with Temperature and Frequency of ϵ'' for i-butyl bromide.

The range of temperature traversed during the setting of the dipoles with respect to a given frequency in liquids and glasses differentiates their rate of intermolecular action from that occurring in crystals in which rotational transitions may occur in a very short or zero temperature interval.^{8, 9, 34} In the latter, the sudden rearrangement of the crystal transition would seem to effect a critical interaction that blocks dipole orientation almost completely. The gradual nature of the process in the glasses, seen in FIGURES 3 and 5, may be further evidence for random molecular arrangement.

³³ Danforth, W. E. Phys. Rev. 38: 1224. 1931.

³⁴ See Figures 7, 8, 10 and Baker, W. O., & Smyth, C. P. Jour. Am. Chem. Soc. 61: 2798. 1939.

Even before the dipole theory, it was realized that the constituent particles of a dielectric did not all have necessarily the same relaxation time.³⁵ This concept has been applied to a variety of dielectrics by many workers, and recently Yager³⁶ has reviewed and extended quantitative estimates of the distribution. Ordinary experimental values represent an average. A rough qualitative idea of the breadth of the distribution issues from a calculation of the ideal maximum value of ϵ'' for a given frequency, as has been done for several compounds by Morgan.³⁷ An observed maximum value obtained from FIGURE 4 or FIGURE 6 will deviate the more from the ideal value, which it would have if all the particles had a single relaxation time, the less uniform are the composite units of the dielectric. Computation of the ideal ϵ''_{max} follows from the observation that the absorption is greatest when $x = 1$ in the Debye expression for ϵ'' .³⁸ That is, $\epsilon'' = x(\epsilon_1 - \epsilon_0)/(1 + x^2)$, where $x = \omega\tau(\epsilon_1 + 2)/(\epsilon_0 + 2)$, in which ϵ_1 = dielectric constant at zero frequency, obtained from graphical extrapolation of low frequency values; ϵ_0 = dielectric constant at infinite frequency, obtained from low temperature values for the crystalline solid; $\omega = 2\pi \times$ frequency in cycles; τ = relaxation time of particles, in seconds. Then,

$$\epsilon''_{max} = \frac{(\epsilon_1 - \epsilon_0)}{\epsilon_0}$$

TABLE I

	<i>f</i> , kc.	ϵ''_{max} Calcd.	ϵ''_{max} Obsd.
i-Butyl bromide.....	50	7.5	5.3
	5	6.7	5.6
	0.5	4.8	5.9
i-Amyl bromide.....	50	5.0	3.8
	5	3.9	4.5
	0.5	3.0	5.0

Seemingly, from TABLE 1, the most uniform relaxation times are characteristic of the temperature region in which ϵ'' is a maximum for $f = 5$ kc. FIGURES 4 and 6 support this inference in showing the 5 kc. dispersion curves to bound narrow areas, with sharp peaks, a further evidence³⁶ for a narrow distribution of relaxation times.

On this basis, isobutyl bromide glass appears to be most homo-

³⁵ Von Schweidler, E. E. Ann. Physik, 24: 711. 1907.

³⁶ Yager, W. A. Physics 7: 434. 1936.

³⁷ Morgan, S. O. Ind. Eng. Chem. 30: 278. 1938.

³⁸ Debye, P. "Polar Molecules." Chemical Catalog Co., New York. 94. 1929.

geneous at about -158° , and isoamyl bromide at about -145.5° , whereas a greater variety of agglomerates exists above and below these temperatures. The molecular aggregates in a glass,^{24, 30} must at least partially dissociate or rearrange before the individual molecules can orient to form a crystal lattice. The ordered crystalline state has a lower free energy than the glass, and hence the latter is forced toward the former by a thermodynamic potential that increases with decreasing temperature. It may be that at just above the temperature of the maximum absorption for 5 kc. in the two glasses, the incipient rearrangement tending toward devitrification

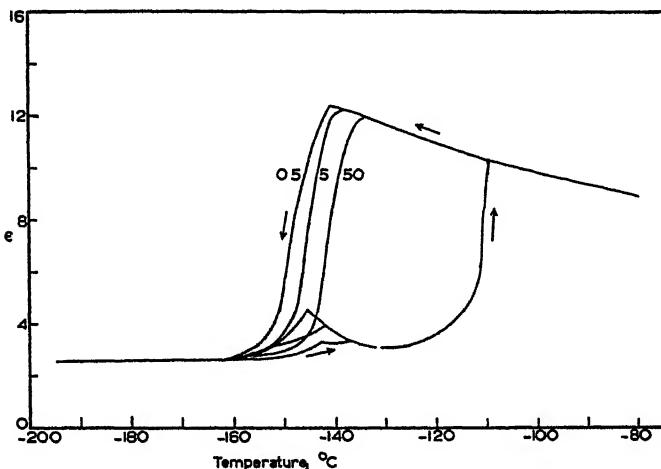


FIGURE 5. Variation with Temperature and Frequency of the Dielectric Constant of 1-amyl bromide.

has produced an effect of maximum particle homogeneity. This does not mean, of course, that the actual orienting units are other than single molecules at any time, but environmental irregularities can cause the apparent size distribution in what is really a rate distribution. The rate concept is made more precise in the next section on the mechanism of orientation in these glasses.

From the above results for the two similar molecules, like relations should obtain for each between the temperature T_r at which the re-ordering enters and the lower temperature T_f of apparent rotational freezing, if temperature difference may be considered as measuring thermodynamic potential. The devitrification tendency could actually be detected at T_f .³⁹ The relations found are

³⁹ Baker, W. O., & Smyth, C. P. Jour. Am. Chem. Soc. 61: 2063. 1939.

$$\text{i-butyl bromide } T_c (-158^\circ) - T_f (-170^\circ) = 12^\circ$$

$$\text{i-amyl bromide } T_c (-145.5^\circ) - T_f (-160^\circ) = 14.5^\circ$$

These temperature intervals are, as expected, about the same, with the larger molecule showing the larger interval.

The portions of FIGURES 3 and 5 indicated by arrows as obtained with rising temperature are now considered. The physical states of i-butyl bromide at -175° and of i-amyl bromide at -165° are those of undercooled liquids far below their normal freezing points (-118° and -112° , respectively). This instability as non-

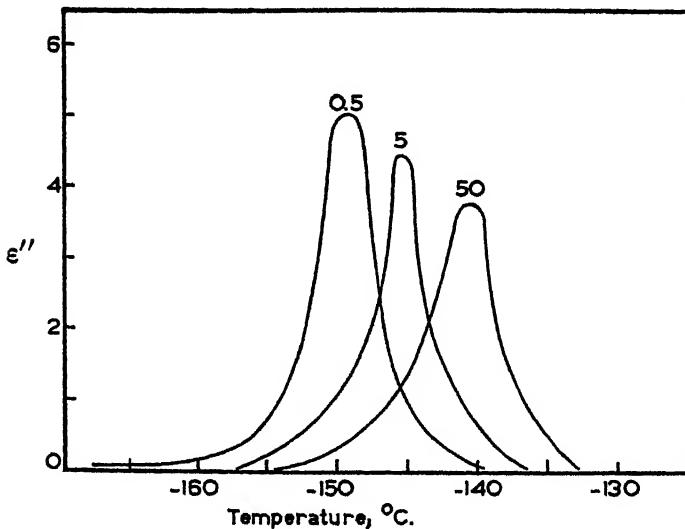


FIGURE 6. Variation with Temperature and Frequency of ϵ'' for i-amyl bromide.

crystalline forms increases with decreasing temperature, but the fruition of the crystallizing tendency is opposed by the increase with lowered temperature of the molecular relaxation times. However, around -170° for i-butyl and -160° for i-amyl bromide, a certain amount of devitrification has occurred, because the dielectric constant-temperature curve from measurements at 0.5 kc. on FIGURE 3 was completely reversible with temperature anywhere down to about -167° , while if the glass was cooled below this point, and then warmed, the dielectric constant values did not return to those required by the reversible curve. Similar results at a higher temperature were found for i-amyl bromide in FIGURE 5. Moreover, the isotropic glasses began to show scattered areas of double refraction

when examined with the polarizing microscope. Because of relaxation effects, the amounts of crystal formation observed at -170° for i-butyl and -160° for i-amyl bromide should depend on the length of time the sample is held at the given temperature after the liquid is cooled. The upper of the two rising temperature curves on FIGURE 3 resulted when the specimen was held near -180° for somewhat more than thirty minutes. Analogous treatment produced the heating curve on FIGURE 5. Apparently there was not time for complete crystallization. Presumably, devitrification had proceeded from many nuclei scattered throughout the glass, but there were still vitreous regions when the measurements with rising temperature were begun. Molecules in the crystal lattice of i-butyl bromide cannot rotate in the applied field, as the lower curve with rising temperature on FIGURE 3 demonstrates, except in the pre-melting region already noted. However, molecules still in the glass distributed among the crystal groups should be able to orient as soon as a sufficiently high temperature is reached. For i-amyl bromide, as seen in FIGURE 5, an increase in dielectric constant is found in about the same temperature region in which it declined during the cooling of the pure glass. For i-butyl bromide, the interval of increase is shifted to somewhat higher temperatures. A speculative reason for the difference is that, in the sample of i-butyl bromide for which the data are graphed, crystallization has progressed so far that the large crystalline groups in a predominantly ordered structure inhibit, by their locally directed force fields, the rotation of the glass molecules dispersed among them. Consistently, the maximum dielectric constant on the warming curve of i-amyl bromide, although it occurs at a lower temperature, is a larger percentage of the true glass value than that for i-butyl bromide. Characteristic glass dispersion entered in the curves for both compounds. Further, the dielectric absorption data³⁹ for both molecules show a maximum with rising temperature near the temperature of the dielectric constant maxima. This also affirms the recovery of partial rotational freedom by those molecules still in the vitreous condition.

Finally, the decrease of the dielectric constant with further temperature rise, seen on the lower curves of FIGURES 3 and 5, occurs because, at the temperature of maximum ϵ , the molecules still in the vitreous state have again acquired sufficiently short relaxation times to crystallize. Actually, crystalline groups already predominate, to act as ordering centers, as seen by an ϵ_{max} on warming of 4.5 instead of 17.3 for the pure glass. As the solid state crystalliza-

tion proceeds, the dielectric constant falls for the refraction value and then later rises on melting.

Probably analogous processes to those explored above occurred in the first production of boron trioxide crystals, recently reported.⁴⁰ An ordinary boron trioxide glass held at 250° for three days, after it had stood at room temperature for a much longer period, displayed considerable crystallization, as verified by X-ray analysis. The melting point of a nearly glass-free mass of boron trioxide crystals (99.6%) was 460–470°. The glass at room temperature, 450° below the normal freezing point, may be regarded as under a large thermodynamic potential. Some few very minute crystal nuclei probably are actually formed. Indeed, such have been postulated for laboratory glass at ordinary temperatures.⁴¹ Then, the temperature increase shortens the relaxation time so that molecules or ionic groups may crystallize extensively.

Results from these and related molecular structures whose behavior has been studied in detail¹⁰ indicate relations of the structure of the liquid state to the tendency toward glass formation as follows:

a. Moderately unsymmetrical, pear-shaped molecules like i-butyl and i-amyl bromide, and similar substituted or branched hydrocarbons can pack randomly in the liquid with an economical use of volume. Hence, although the molecules are on the whole disordered, their efficient packing allows persistence of the liquid arrangement far below the normal crystallization temperature. Such substances should vitrify. Additional confirmation for this is the observation⁴² that two of the appropriately branched heptanes, 3-methylhexane and 2, 3-dimethylpentane, unlike the others, failed to crystallize even at 70° K and the recent report that various other branched hydrocarbons, unlike the linear members, solidified to glasses.⁴³

b. Long-chain compounds, although they often form viscous liquids, do not generally supercool.^{12, 44} It is probable that the rod-like molecules are already locally, at least, well aligned in the liquid. Random array would waste a relatively large volume, and is thus energetically unlikely. Indeed, X-ray data indicate long axis parallelism even in n-heptane.^{45, 46} Hence, when long chain

⁴⁰ McCulloch, L. *Jour. Am. Chem. Soc.* **59**: 2650. 1937.

⁴¹ Warren, B. E., & Briscoe, J. *Jour. Am. Ceram. Soc.* **21**: 259. 1938.

⁴² Huffman, H. M., Parks, G. S., & Thomas, S. B. *Jour. Am. Chem. Soc.* **52**: 3241. 1930.

⁴³ Smittenberg, J., Hoog, H., & Henkes, R. A. *Jour. Am. Chem. Soc.* **60**: 17. 1938.

⁴⁴ See diagrams in Smith, C. P. *Ann. Reports.* **35**: 251. 1938.

⁴⁵ Pierce, W. C. *Jour. Chem. Phys.* **3**: 252. 1935.

⁴⁶ Katzoff, S. *Jour. Chem. Phys.* **2**: 841. 1934.

molecules are cooled to the freezing point, they simply set into the lattice, without marked supercooling.

c. Very symmetrical, nearly spherical molecules which form cubic lattices, and especially those which are known to possess a rotational freedom in the crystal comparable to that in the liquid, commonly do not supercool. They may be considered to compose a liquid structure of close-packed spheres, which can set into the cubic lattice with little rearrangement and probably no reorientation, and thus they have little tendency toward vitrification.

d. As expected and noted in the cases of i-propyl and n-amyl bromides previously, molecules of symmetry intermediate between pear-shaped and rod-like show degrees of supercooling varying with their particular form, and very rarely leading to true glass formation.

Extensive intermolecular bonding in a liquid may so effectively alter the contours of the constituent particles as to require modification of the above conclusions from molecular data alone.

RELAXATION MECHANISM IN GLASSES

The dispersion data provided over a temperature and frequency range for the i-butyl and i-amyl bromide glasses³⁹ are now considered in terms of the theories of dielectric relaxation and its accompanying molecular motion. The latter will be related to the current concepts of viscous flow. Dielectric loss related to the polar structure of matter has been comprehensively revised by Mueller.⁴⁷

Debye⁴⁸ applied Stokes' expression for the rotation of a sphere in a viscous medium to obtain a relation between the relaxation time, τ , of the molecules, and their radius, a , in which

$$\tau = \frac{4\pi\eta a^3}{kT}$$

where η is the inner friction constant, k is the Boltzmann constant, and T is the absolute temperature. This equation is occasionally moderately well obeyed by dilute solutions of polar molecules, when η is taken as the viscosity of the solvent, but requires revision when there is pronounced intermolecular action, as in a pure polar substance. This Debye⁴⁹ and others have recognized. Then, η can no longer be a simple macroscopic viscosity, and we shall consider τ in terms of molec-

⁴⁷ Mueller, H. *Ergebnisse der Exakten Naturwissenschaften*. Julius Springer. Berlin 17: 164-228. 1938.

⁴⁸ Debye, P. "Polar Molecules." Chemical Catalog Co., New York. 85. 1931.

⁴⁹ Debye, P., & Ramm, W. *Ann. Physik*. 28: 28. 1937.

ular processes. For comparison with the latter treatment, however, the classical relaxation times of the two bromides have been calculated. By a graphical extrapolation ϵ_1 , the dielectric constant at zero frequency is evaluated and ϵ_0 , that at infinite frequency, is found from the low temperature solid value. The calculation of τ proceeds from Debye's equation,⁵⁰

$$\omega\tau = \frac{\epsilon_0 + 2}{\epsilon_1 + 2} \cdot [\frac{1}{2}\{y + (4 + y^2)^{\frac{1}{2}}\}]^{\frac{1}{2}}$$

where $\omega = 2\pi f$, f = frequency in cycles per second, $y = (\epsilon_1 - \epsilon_0)/(\epsilon_1 + \epsilon_0)$. The relaxation times, in each case at a temperature corresponding to maximum dispersion for 50 kc., are:

i-butyl bromide, $\tau = 0.9 \times 10^{-6}$ sec. at 118.6° K.

i-amyl bromide, $\tau = 1.2 \times 10^{-6}$ sec. at 131.1° K.

Rough estimates of kinetic theory radii from Sutherland's constant data (Landolt-Bornstein "Tabellen") may be used to calculate η values: i-butyl bromide, for $a^3 = 16.64 \times 10^{-24}$ cm³, $\eta = 0.68 \times 10^2$ poise at 118.6° K; i-amyl bromide, for $a^3 = 20.91 \times 10^{-24}$ cm³, $\eta = 0.79 \times 10^2$ poise at 131.1° K. Lillie⁵¹ showed that there is no discontinuity in η at the vitrification point, while Eyring⁵² has justified theoretically a linear relation between η and $1/T$ which allows extrapolation of the viscosity data of Thorpe and Rodger⁵³ for i-butyl bromide to the temperature at which the relaxation time was measured. Although the value thus obtained, $\eta = 0.90$ poise at 118.6° K, is very approximate, it demonstrates, by being 100 times smaller than the doubtfully significant "classical" value, the inadequacy of the simple macroscopic viscosity in specifying the rotational hindrance in a pure polar medium. Actually, also, the true inner friction constant at the temperatures involved would represent a large plastic flow component.

A method for considering the relaxation phenomenon as a molecular process has been suggested by Eyring,^{52, 53a} and further developed by Frank,⁵⁴ who made calculations for ice, some solid solutions in paraffin wax, and for a composite insulator, Permitol. We pro-

⁵⁰ Debye, P. "Polar Molecules." Chemical Catalog Co., New York. 94. 1929.

⁵¹ Lillie, H. R. Jour. Am. Ceram. Soc. 16: 619. 1933.

⁵² Eyring, H. Jour. Chem. Physics. 4: 283. 1936. See also Andrade, E. N. De G. Nature. 125: 580. 1930. Sheppard, S. E., & Houck, E. C. Jour. Rheology. 1: 349. 1930.

⁵³ See Landolt-Bornstein. Tabellen. Julius Springer. Berlin. 5th ed. 1. 1935.

^{53a} Stearn, A. E., & Eyring, H. Jour. Chem. Phys. 5: 113. 1937.

⁵⁴ Frank, F. C. Trans. Faraday Soc. 32: 1634. 1936.

pose to extend this treatment to the simple homogeneous bromide glasses. The quantal basis for the theory of absolute reaction rates⁵⁵ gives an expression for the velocity of decomposition into products of an activated complex formed at the top of a potential hump. For a unimolecular reaction, this is effectively the surmounting of a potential barrier, which may readily be considered as that imposed by its neighbors to restrict the rotation of a given molecule. The standard Arrhenius form yields an activation energy for the process, since $\log \tau$ is found to be practically a linear function of T , and $1/\tau = k' = Ce^{-A/RT}$, where k' = specific reaction rate, A = activation energy, C = "steric factor," later to be explicitly defined. The required τ is most accurately found by regarding it as the mean relaxation time for the temperature of maximum dielectric loss. That is, considering the decay of orientation as an exponential function of time, Debye⁵⁰ has derived the relations

$$\epsilon' = \epsilon_0 + \frac{\epsilon_1 - \epsilon_0}{1 + x^2}$$

$$\epsilon'' = \frac{\epsilon_1 - \epsilon_0}{1 + x^2} x$$

in which the quantities have been already defined.

ϵ'' is the loss component of the dielectric constant, from $\epsilon = \epsilon' - i\epsilon''$. Clearly, ϵ'' reaches a maximum when $x = 1$, and the temperatures of these maxima are available from the experimental results independently of the absolute values of ϵ'' , since its graphs against temperature are quite symmetrical, as shown in FIGURES 4 and 6. We now may write

$$-\frac{d \ln \tau}{dT} = \frac{A}{RT^2}, \text{ or, since } \omega = 2\pi f,$$

and $\tau = 1/\omega$ at the critical frequency,

$$\frac{d \ln f}{dT} = \frac{A}{RT^2}.$$

Thus, $\frac{A}{R}$ values result from the slope of the straight line obtained by plotting $\ln f$ against $1/T$, where T is the temperature of maximum ϵ'' for the given frequency f . Further, a mean T of ϵ''_{max} may be found, and from it a mean f , which may be then used to calculate the specific rate, a mean $1/\tau$, from

⁵⁵ Myring, H. Chem. Rev. 17: 65. 1935.

$$1/\tau = \frac{\epsilon_1 + 2}{\epsilon_0 + 2} \cdot 2\pi j.$$

The fundamental equation $k' = \kappa \frac{kT}{h} e^{-\frac{\Delta F^\ddagger}{RT}}$, where κ = transmission coefficient, taken as 1, k = Boltzmann's constant, h = Planck's constant, and ΔF^\ddagger = the free energy of activation, may now be used to calculate ΔF^\ddagger . The equation may be recast as $k' = \kappa \frac{kT}{h} e^{\frac{\Delta S^\ddagger}{R}} e^{-\frac{\Delta H^\ddagger}{RT}}$, where ΔS^\ddagger is the activation entropy and ΔH^\ddagger is the heat of activation taken as the activation energy A for the rotation process. The factor C thus includes the terms $\kappa \frac{kt}{h} e^{\frac{\Delta S^\ddagger}{R}}$. The results of TABLE 2 elucidate the relaxation mechanism.

TABLE 2

	T mean °K	f mean	A (Cal./mole)	C	ΔF^\ddagger (Cal./Mole)	ΔS^\ddagger (E. I./Mole)
i-Butyl bromide	115.6	5.6×10^3	23,100	7.86×10^{40}	3796	167.1
i-Amyl bromide	128.0	5.0×10^3	18,000	5.21×10^{46}	4425	106.1

It is evident from the large entropy terms that many molecules surrounding a given one able to orient in the field must cooperate in permitting its movement from one potential valley to another, not necessarily equivalent, one. Probably they undergo, by thermal fluctuations, slight rearrangements, a very definite combination of these being essential for the single rotation. In the liquid state, this entropy factor is of vastly diminished importance, because the energy differences of various configurations are diminished by the lower density and higher heat content. The demonstrated necessity for this cooperative effect in "rigid" phases helps to explain the slowness of diffusion and flow in glass, and the like; there will always be a competition between the heat and entropy terms.

Although the additional methylene group in i-amyl bromide somewhat dilutes the dipoles, this alone is insufficient to account either for the lower activation energy for rotation, 18 kcals, or for the reduced maximum dielectric constant, 12.3, of the larger molecule as compared to values of 23 kcal and 17.3, respectively, for i-butyl bromide. However, this difference would be rationalized if the elongated pear shape of the i-amyl as contrasted to the compressed pear shape of the i-butyl compound had introduced sufficient anisotropy to make the former molecule rotate preferably only about its long

axis. Independent evidence that this slight structural change has indeed produced such an effect, and, conversely, support from the dielectric measurements for the rotating couple concept of viscous flow^{52, 56} arise from inspection of the activation energies of viscosity for these compounds. The latter, listed in TABLE 3, were calculated from the data of Thorpe and Rodger.⁵³ According to the Eyring theory of viscosity, the flow process occurs in a fashion requiring the least use of extra volume, for to make a hole the size of a molecule requires the energy of its vaporization. For nearly spherical, or quite symmetrical, structures, no preferred orientations of the molecules forming the rotating couple are possible, and the ratio of the energy of vaporization to the activation energy for viscous flow, $\Delta E_{vap}/\Delta E_{vis}$, is generally about 3.⁵⁶ However, rod-like molecules will align by long axes, and then the ratio $\Delta E_{vap}/\Delta E_{vis}$ is 4 or even more. If, then, i-amyl bromide is tending to behave as an extended molecule, while i-butyl bromide is more compact, the ΔE_{vis} values should be in the same order as, although of quite different magnitudes from, the activation energies for orientation. Table III verifies this, and further exhibits the expected difference between the $\Delta E_{vap}/\Delta E_{vis}$ ratios, 4.3 for the longer molecule and 3.5 for the more compressed.

TABLE 3

	ΔE_{vis} (cal/mole)	A (cal/mole)	$\frac{\Delta E_{vap}}{\Delta E_{vis}}$ (cal/mole)
i-Butyl bromide.....	1990	23,100	3.5
i-Amyl bromide.....	1760	18,000	4.3

Thus it is possible to connect the mechanisms of the kinetic processes of dielectric relaxation and viscous flow, and it becomes simultaneously evident that the constants involved, such as η , are not interchangeable. Such a plan of deciding about the mode of orientation processes should prove general.

It is observed in connection with the swarm theory of liquid crystals⁵⁷ that the relaxation time of para-azoxyanisole, as detected by the region of maximum dispersion for a given frequency, varies rapidly with temperature. However, the experimental viscosity shows a normal low temperature coefficient. Also, optical studies

⁵⁶ Ewell, R. H., & Eyring, H. Jour. Chem. Phys. 5: 726. 1937.⁵⁷ Ornstein, L. S., & Kast, W. Trans. Faraday Soc. 29: 931. 1933.

reveal that the groups are altered in size only gradually by temperature change. Such relaxation time-viscosity phenomena, anomalous in the classical theory, are perhaps understood from the present evidence that the usual viscous flow process requires only a fraction of the activation energy involved in significant measurable orientation in a field. Hence, as found, the temperature coefficient of the relaxation rate would exceed that of the viscosity change.

The viscosity of liquid crystalline systems is also found to show a maximum with increasing temperature in the anisotropic region,⁵⁸ instead of the uniform decrease of normal liquids. Too, the Hagen-Poiseuille law is not obeyed, and liquid crystal mixtures possess a disproportionately high viscosity for low pressures.⁵⁸ Ostwald has termed this a "structure viscosity" effect without further discussion. The structure sensitivity is quantitatively illustrated in the bromides here considered. If the difference of a methylene group so alters the flow orientation, as indicated, it is understandable that liquid crystalline groups may, by relatively small changes in shape, engage in a variety of flow processes. The changes in shape are effected by temperature, and perhaps changes in group orientation result from capillary streaming, hence the observed anomalies. The dielectric results have shown how sensitive potentials involving rotational interaction are to the shape of the interacting units.

Oncley and Williams⁵⁹ have noted the deficiencies of the macroscopic viscosity in relaxation time calculations in reporting that the measured viscosity of a solution might be increased a thousand-fold without changing significantly the value of the inner friction constant necessary in the Debye formula. The viscosity was varied by using mixtures of hydrocarbon oils or waxy compounds. From the present scheme of regarding orientation as the surmounting by the polar solute molecules of potential barriers created, at least in moderately dilute solutions, chiefly by the solvent molecules, lengthening of the latter would presumably not alter the average environment of the solute. Hence, just the results of Oncley and Williams would be expected, as during the rotation of the polar molecule in the field in contrast to the viscous flow of the solution, the solvent molecules need not themselves rotate.

⁵⁸ Ostwald, W. *Trans. Faraday Soc.* **29**: 1002. 1933.

⁵⁹ Oncley, J. L., & Williams, J. W. *Phys. Rev.* **43**: 341. 1933.

SPHERICALLY SYMMETRICAL MOLECULES

Finally, in this series, we consider the t-butyl halides, whose groups are tetrahedrally disposed⁶⁰ about a central carbon atom, and which, despite their considerable permanent polarity, are highly symmetrical. They behave as spherical molecules in viscous flow. They may also be regarded in their crystal and dielectric behaviors as expanded analogues of the hydrogen halides, which have been previously studied.⁶¹⁻⁶⁴ Dielectric, optical and some thermal examina-

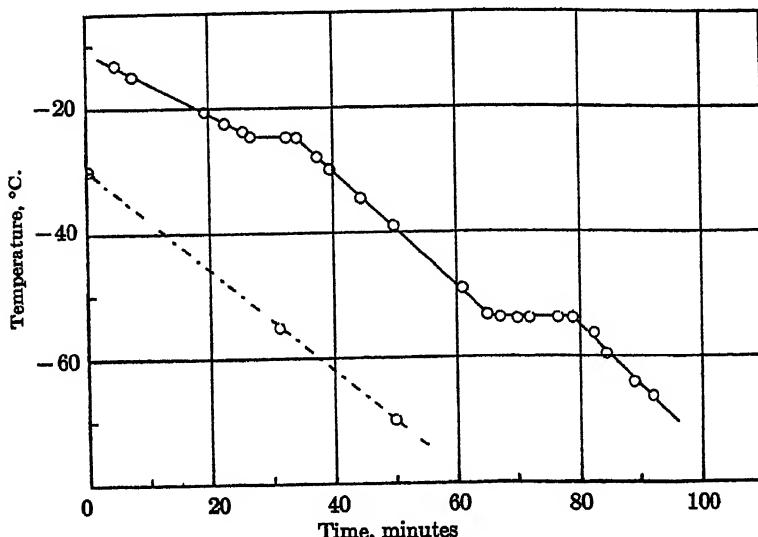


FIGURE 7. Temperature time curve (cooling) for t-butyl chloride.

tions of t-butyl chloride, bromide and iodide demonstrated rotational transitions in each.⁶⁵

The cooling curve for t-butyl chloride in FIGURE 7 reveals a previously unreported thermal transition at -53.8° , whose latent heat considerably exceeds that of fusion. The temperature variation of the dielectric constant resembles that of hydrogen chloride, as seen from FIGURE 8. The sections of the curves obtained on cooling, how-

⁶⁰ For exact values of the interatomic distances in t-butyl chloride, see Beach, J. V., & Stevenson, D. P. *Jour. Am. Chem. Soc.* 60: 475. 1938.

⁶¹ Smyth, C. P., & Hitchcock, C. S. *Jour. Am. Chem. Soc.* 55: 1830. 1933.

⁶² Hettner, G. & E., & Pohlman, R. *Zeit. Physik.* 108: 45. 1937.

⁶³ Damköhler, G. *Ann. Physik.* 31: 76. 1938.

⁶⁴ Kruis, A., & Kaischew, R. *Zeit. physikal. Chem. B* 41: 427. 1939.

⁶⁵ Baker, W. O., & Smyth, C. P. *Jour. Am. Chem. Soc.* 61: 2798. 1939.

ever, slope downward with decreasing temperature while detailed results on the cooling behavior of hydrogen chloride are not available for comparison. The highest temperature point in the liquid on FIGURE 8 was obtained independently in a dipole moment study.⁶⁶ On solidification the dielectric constant rises because of increase in density, shows a low sharp peak, presumably from a small Maxwell-Wagner interfacial polarization, and then gradually declines with a small but definite dispersion. At the transition temperature, the

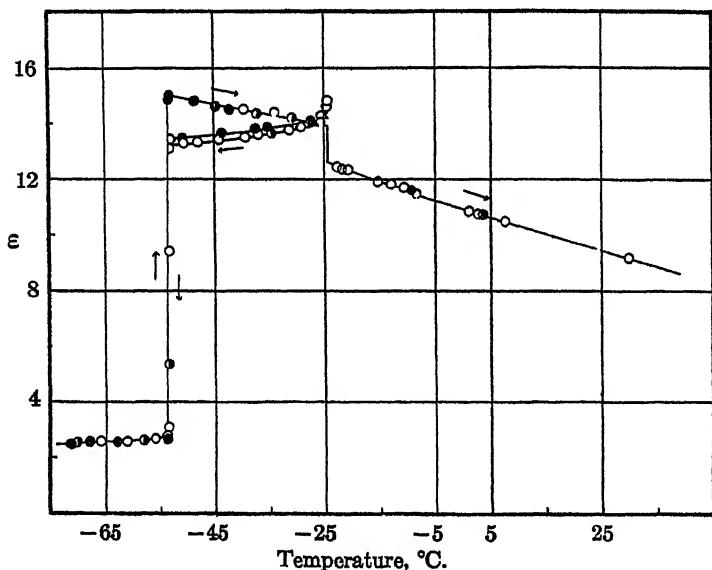


FIGURE 8. Temperature Dependence of the Dielectric Constant of *t*-butyl chloride. (Hollow circles represent values at 50 kc, half-filled circles values at 5 kc., and filled circles values at 0.5 kc.)

dielectric constant drops sharply to nearly the square of the refractive index, with disappearance of dispersion. The polarizing microscope revealed the transition as an enantiomorphic transformation from a high temperature cubic to a low temperature form of lower symmetry. The abrupt lattice change into an anisotropic packing arrangement would follow from a cooperative interlocking of the dipoles when they had reached a sufficiently low temperature to lose most of their vibrational energy. Such a sharp interaction is implied by FIGURE 9, where the curve denoting the lower ϵ'' values in the region of -53.8° rises steeply to a maximum just at the transition

⁶⁶ Smyth, C. P., & Dornte, R. W. Jour. Am. Chem. Soc. 53: 545. 1931.

temperature. Hence a rapid and widespread interference to orientation is introduced by a loss maximum so narrow that it is only a straight line for 50 kc. at the peak. Then the dipoles are immobilized in the low temperature crystalline form.

The dielectric constant curve obtained with rising temperature, shown by arrows on FIGURE 8, exhibits a sharp ascent at the transition, which is unattended by thermal hysteresis, to values which decline on further warming, as expected from the Debye relation.

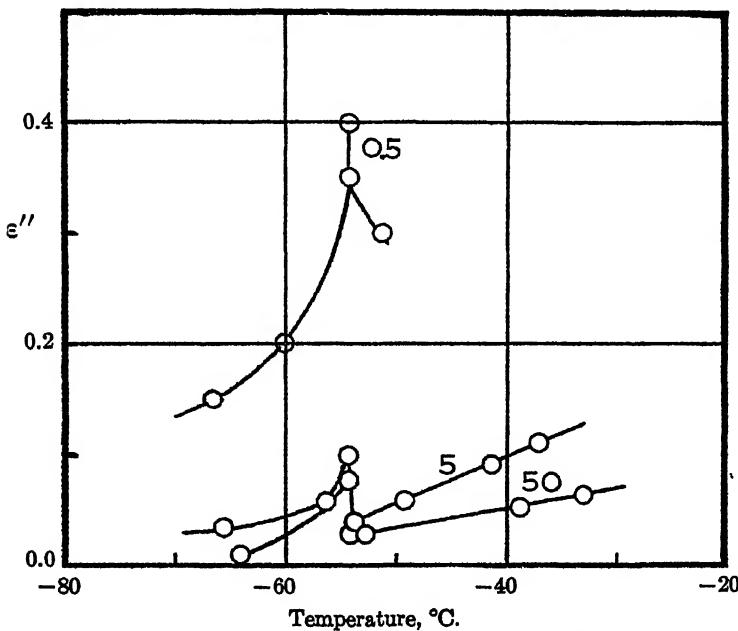


FIGURE 9. Temperature Dependence of the Imaginary Part ϵ'' of the Dielectric Constant of *t*-butyl chloride.

If the cooperative loosening of the molecules at the transition on warming is great enough to yield immediately the maximum orientational freedom in the cubic lattice, whereas, on cooling, a decrease in the molecular disorder, perhaps beginning immediately after solidification, gradually reduces this freedom, the observed difference in the warming and cooling curves would be accounted for. Microscopic examination revealed qualitatively a considerable density increase on cooling through the transition.

FIGURE 10 contains the thermal arrests marking the two polymorphic transitions found for *t*-butyl bromide. The higher tempera-

ture transition, henceforth called I, is attended by supercooling but comparison of the heating and cooling curves, together with the dielectric results, suggest that it occurs without thermal hysteresis at -41.4° . The lower transition II shows no supercooling but occurs abruptly at -65.8° on cooling and at -64.2° on warming. Evidently, from FIGURE 10 the heat of fusion is less than the heat of transition II, and much less than the sum of the heats of the two transitions.

An enhanced dispersion in the same frequency range results from replacement of a chlorine atom by the larger, more polarizable bro-

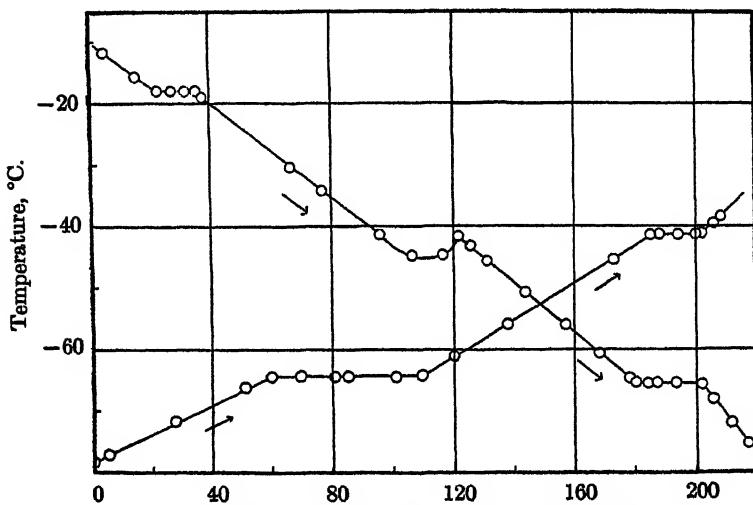


FIGURE 10. Temperature time Curves for t-butyl bromide.

mine, as is evident from FIGURE 11. In some parts of the graphed data points at the three frequencies would have been superposed, so but one point was indicated, for example, the point immediately above transition II, on cooling. The data for the lower frequencies begin somewhat below the freezing point, for, just after solidification bridge balancing became uncertain and large apparent dielectric constant values were obtained probably because of interfacial polarization. This effect has been found by Yager and Morgan⁶⁷ for several camphor derivatives, and by White and Morgan⁶⁸ for ethylene cyanide, all of which possess rotational freedom in the solid. The effect here does not exceed a maximum apparent specific con-

⁶⁷ Yager, W. A., & Morgan, S. O. Jour. Am. Chem. Soc. 57: 2071. 1935.

⁶⁸ White, A. H., & Morgan, S. O. Jour. Chem. Phys. 5: 655. 1937.

ductance of 5×10^{-8} ohm $^{-1}$ cm $^{-1}$, although they report values greater than 10 $^{-6}$ ohm $^{-1}$ cm $^{-1}$. Values obtained at 520 kilocycles with a heterodyne beat apparatus are shown through the melting point by the tagged circles.

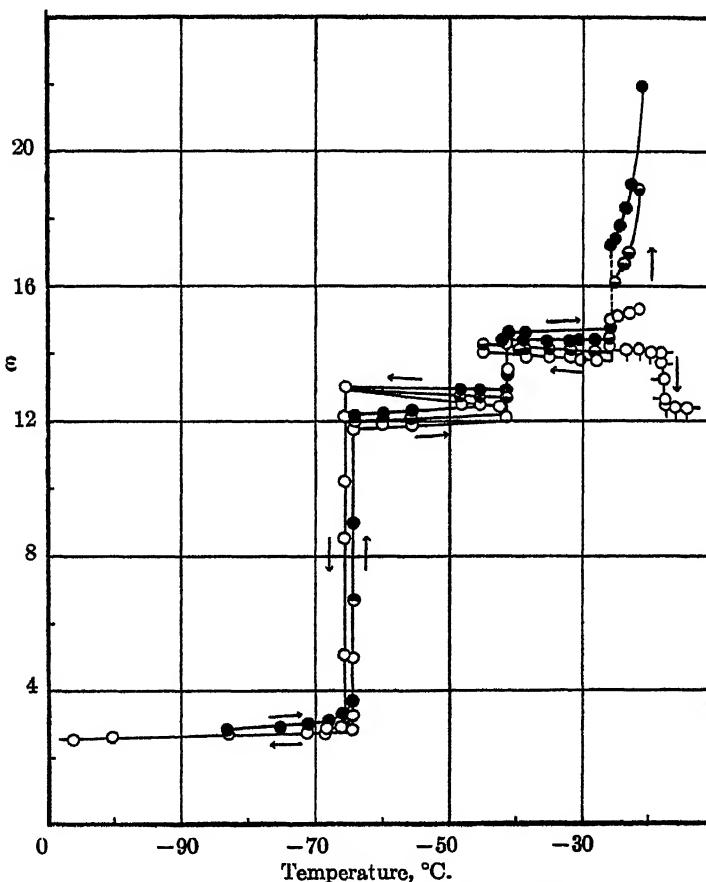


FIGURE 11 Temperature Dependence of the Dielectric Constant of t-butyl bromide. (Hollow circles represent values at 50 kc., half-filled circles values at 5 kc., and filled circles values at 0.5 kc.)

The supercooling at transition I is manifest on the dielectric graph by an exact continuation of the curves for the highest temperature solid form some 3° below -41.4°. Suddenly, the values revert back on just these lines, as far as could be determined experimentally, and the dielectric constant drops sharply at the transition to a new set of curves, which show diminishing dispersion with decreasing tempera-

ture. At -65.8° , a transition like that in t-butyl chloride occurs. However, below this point for a short temperature interval there is still dispersion, apparent in the data but too slight to be evident in FIGURE 11. As with the chloride, a sharp rise in ϵ'' marks the interaction introduced by transition II.

On warming, a slight rotational pre-melting, consonant with the suggestion of cooperative lattice loosening with rising temperature noted for t-butyl chloride, is followed by the onset of rotation, now at -64.2° . This slight rotational freedom below a transition seems comparable to the pre-melting previously noted for many molecular lattices.¹² The latter behavior has been confirmed by X-ray and thermal analysis of highly purified hydrocarbons,²⁰ and has been given theoretical interpretation and justification by Frank's recent order-disorder treatment of melting.⁶⁹ It would seem that in a treatment of arrangement in terms of rotational disorder and of the corresponding lattice energy change dependent on the degree of such disorder, a transition state just below the observed transition temperature like that proposed by Frank just below the melting point would account for the experimental data.

The apparent 1.6° hysteresis in the temperatures of transition resembles the 1.1° difference in the case of the analogous transition in hydrogen sulfide,⁷⁰ the longer interval in ethylene cyanide,⁶⁸ and the less sharply defined hystereses in camphor and its derivatives.⁶⁷ The same hystereses are found in thermal measurements, and were thus discovered at the rotational transitions of methane and deuterated methane.⁷¹

Pressure studies by Clusius and Weigand⁷² on the hydrogen sulfide transition disclose that it is accompanied by a volume change of about 7.6%, and by a mean lattice separation change so small that it might escape detection in X-ray photographs. However, easily obtainable pressures with hydrogen gas broadened the hysteresis range, and the hysteresis area was found to be smaller in deuterium sulfide. Hence, apparent hysteresis in these transitions would seem to be sensitively affected by changes in structure and environment. The apparent temperature of the monotropic transition in ethyl stearate was found to depend on the temperature gradient to which the sample was exposed.¹² This result was confirmed for other

⁶⁹ Frank, F. C. Proc. Roy. Soc. London. 170: 182. 1939.

⁷⁰ Smyth, C. P., & Hitchcock, C. S. Jour. Am. Chem. Soc. 56: 1084. 1934.

⁷¹ Krnis, H. A., Popp, L., & Clusius, K. Zeit. Elektrochem. 43: 664. 1937.

⁷² Clusius, K., & Weigand, K. Zeit. Elektrochem. 44: 674. 1938.

similar long chain compounds.⁷³ It provides evidence for the idea that critical fluctuations in ΔF and ΔS must occur for the progression of spontaneous solid transformations. This concept has been elaborated by Ubbelohde.⁷⁴ The activation entropies for rotation of even simple molecules in the solid state have been found very high.⁷⁵ Therefore, it is possible that the t-butyl bromide hysteresis occurs because, in order that transformation may be induced, the substance must be cooled slightly below the true equilibrium point to produce an enhanced thermodynamic potential. Then, the material may lose heat to the environment so readily that the temperature of the sample never rises to the equilibrium value, despite the relatively large latent heat. This agrees with the observation that the sample could be cooled under a larger gradient and made to lose heat so rapidly that the transition appeared to occur at -67.2° . However, the process at -65.8 as shown by FIGURES 10 and 11, is not an ordinary supercooling like that at transition I.

A macroscopic potential, or tension, in the crystal mass produced by the transition has been suggested by Frank^{76, 77} as the possible cause of hysteresis. The rate of temperature change as the transition is approached should strongly influence the hysteresis, it would seem. Further limitations on this idea have been noted by Eucken.⁷⁸

An explanation applicable to certain examples of hysteresis is suggested by the dielectric and thermal results on n-amyl bromide,¹⁸ and by the X-ray studies of Wallerant on camphor.⁷⁹ The latter found that besides the cubic form existing above the transition, three less symmetrical classes occur at successively lower temperatures. Monotropic transitions below a rotational transition would produce the same effect on it as that of the n-amyl bromide transition in raising the melting point several degrees above the freezing point. A careful thermal and X-ray analysis of other substances showing hysteresis might establish such low temperature polymorphism.

Examination with the polarizing microscope showed t-butyl bromide to form a translucent, isotropic, definitely crystalline lattice on solidification, which was transformed enantiotropically at transition II to or from an opaque, birefringent, anisotropic lattice. Transi-

⁷³ Vold, R. D., & Vold, M. J. *Jour. Am. Chem. Soc.* **61**: 808. 1939.

⁷⁴ Ubbelohde, A. R. *Trans. Faraday Soc.* **33**: 1203. 1937.

⁷⁵ Baker, W. O., & Smyth, C. P. *Jour. Chem. Phys.* **7**: 574. 1939.

⁷⁶ Frank, F. C., & Wurtz, K. *Naturwiss.* **26**: 687. 1938.

⁷⁷ Frank, F. C. *Zeit. Elektrochem.* **45**: 150. 1939.

⁷⁸ Eucken, A. *Zeit. Elektrochem.* **45**: 150. 1939.

⁷⁹ Wallerant, F. *Compt. rend.* **158**: 507. 1914.

tion I was accompanied by no change in the cubic form, but, apparently, by a contraction on cooling and an expansion on warming. This is indicated by the migration of the Becke line⁸⁰ across the sample as it transforms. In a homogeneous medium, differences in refractive index, detected under the polarizing microscope by brightness variations with a constant source, result from density differences. At the vertical junction between a more dense phase *B* and a less dense one *A*, $n_B > n_A$, where n is the refractive index. The rays which pass from *A* to the denser medium *B* are bent toward the normal to the junction within *B*. Rays passing from *B* toward the junction which impinge within the critical angle are totally reflected and emerge on the *B* side. The result of these combined effects is to illuminate more brightly the *B* side of the junction. Here the junction is the interface between the transformed and untransformed material, and at the transition point a darkish line is seen to move across the sample. It is concluded that I is an enantiotropical lattice change shown by FIGURE 10 to possess a latent heat and accompanied by a density shift which does not destroy the cubic arrangement.

The supercooling at transition I, shown on FIGURES 10 and 11, is like that frequently observed at solid-liquid transitions. Thermal measurements on phosphine disclosed a low temperature rotational transition and an intermediate one,⁸¹ analogous to those of t-butyl bromide. The intermediate transition in phosphine also shows supercooling, whereas the lower one, which presumably marks the end of librational freedom, does not, in agreement with the t-butyl bromide results. The dielectric data provide a possible interpretation of the general property of supercooling in some polymorphic transitions, and its absence in others. FIGURE 10 indicates a much smaller latent heat for transition I than for transition II. The difference between the entropy changes at the two points is even larger. A metastable condition of the solid phase which is stable above the temperature of transition I induced by cooling it somewhat below the transition temperature will be under a relatively small thermodynamic potential tending to force the transformation. Further, FIGURE 11 indicates by the relatively small decrease in dielectric constant at transition I only a slight loss in orientational freedom of the dipoles. Hence, on cooling, the thermal agitation of the molecules which, at a given instant, have undergone the transition will make them poor ordering

⁸⁰ See Hartshorne, N. H., & Stuart, A. "Crystals and the Polarizing Microscope." Edward Arnold & Co. London. 1934.

⁸¹ Clausius, K., & Frank, F. C. Zeit. Physikal. Chem. B34: 405. 1936.

centers for the cooperative formation of the lower temperature stable form. On the other hand, transition II, and, presumably also, as a dielectric study should reveal, the lowest transition in phosphine, mark the loss of most of the rotational molecular freedom. The molecules cease to exhibit nearly spherically symmetrical force fields, and become so definitely oriented that a rearrangement occurs, resulting in the formation of an anisotropic lattice. These effects are evidenced by the loss of orientational polarization seen in FIGURE 11, and the large latent heat of the transition indicated in FIGURE 10. The same causes account for the absence of supercooling at the transition in t-butyl chloride. Such considerations would seem

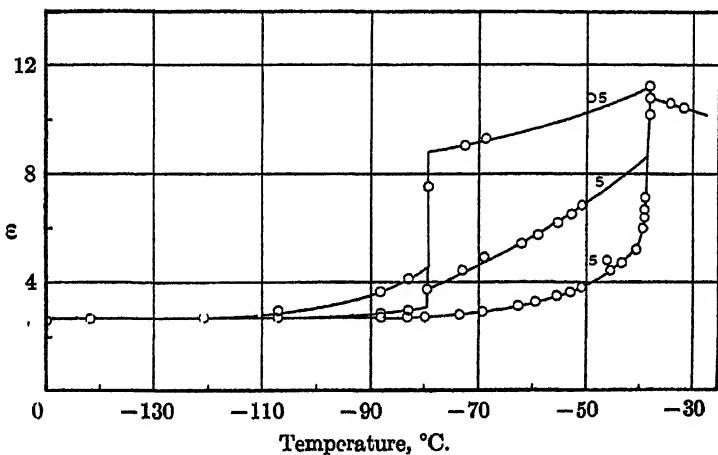


FIGURE 12. Temperature Dependence of the Dielectric Constant of t-butyl iodide.

generally adaptable to thermal studies of polymorphism in molecular and some ionic lattices. They are consonant with other ideas on the importance of orientation in supercooling, noted previously.

FIGURE 12 indicates the expected result that the least symmetrical, most polarizable molecule of the series, $t\text{-C}_4\text{H}_9\text{I}$, interacts in the solid so strongly that it can rotate only with difficulty above the transition point -79.4° . This hindrance effects the pronounced dispersion seen on the graph, but the dielectric loss, proportional to the tabulated apparent specific conductance, contains a considerable d.c. conductance contribution from impurities which could not be removed.⁶⁵ The critical frequency of the rotating dipoles is probably near 50 kc., as the dielectric constant values for this frequency rise from nearly the square of the refractive index, only shortly be-

low the melting point -38.2° . Sufficiently high frequencies might convert the chloride and bromide curves to ones resembling the iodide. Examination with the polarizing microscope revealed an apparently anisotropic lattice above the transition, which was replaced by one of much stronger birefringence below it. Suspected impurities make uncertain a definite assignment of non-cubic class to the high temperature form.

The present results show the variation in intermolecular action of nearly spherical molecules produced when atoms of gradually increasing size and polarizability are successively substituted as one of four groups about a central atom, the other three groups remaining unchanged. As noted previously, the orientational interaction is found to be very sensitive to structure. A further method of studying the possible coupling is to dilute the dipoles in the solid state, preferably by solid solution formation with a similar, non-polar substance. A preliminary microscopic examination of a 50% by volume mixture of t-butyl chloride and carbon tetrachloride revealed that, as would be expected, the transition was made less sharp, and was displaced to lower temperatures, and that the whole solid formed an isotropic high temperature system, which would be anticipated from the known high temperature cubic form of carbon tetrachloride. However, partial melting was found to occur sharply and uniformly, leaving a skeleton-like isotropic lattice which itself melted sharply a few degrees higher. Dielectric measurements on such a system might show a complex behavior because of the coexistence of phases.

The discovery of rotational freedom in the solid forms of these simple molecules suggests experiments to discover its effect on other physical properties. That is, the form of the solid above the transition might differ significantly from that below it in temperature coefficient of thermal conductivity, diffusion, solution rate and sublimation vapor pressure. If these processes were thus connected with rotation or non-rotation, information on their mechanism would be gained.

MOLECULAR FREEDOM AND FUSION

We shall now examine briefly some of the thermodynamic consequences of the solid state behavior of comparable molecules reviewed in this report. The dielectric studies have provided a mechanical estimate of the order and mobility obtaining in molecular lattices or vitreous aggregates of polar molecules. Such studies have indicated marked differences among very similar, in some cases

isomeric, molecules. It may then be expected that the entire energetics of the given solid will be affected, and the striking and fundamental property of the melting point was chosen for inquiry. We regard first the melting properties of the simple alkyl halides, as an approach to subsequent generalizations. The mono-halides, and especially the bromides, are chiefly considered, because they have closely similar polarizabilities and permanent moments, obey Trouton's Rule well as shown by available data,⁵³ and exhibit no extensive liquid association. They should thus have comparable behavior in condensed phases. Also, the dimensions of the molecules should be little altered by interchanges of methyl and bromo groups, since the same method of measurement yields 2.0 and 1.9 Å, respectively for their radii.⁵² Even a chlorine atom with radius of 1.65 Å approximates the methyl group in size. Hence, it is not surprising that the relationships noted below apply to the isomeric hydrocarbons as well as to the isomeric halides.

Writers on the theories of fusion and of liquids have considered the thermodynamic and statistical changes accompanying melting.^{53, 54, 55} Hirschfelder, Stevenson and Eyring explained the low entropies of fusion of substances, which gained non-translational degrees of freedom in the solid state (as by rotational transitions) as arising from the communal sharing of free volume possible on liquefaction. They have shown that if complete communal sharing immediately occurs,⁵⁶ the entropy of fusion ΔS_f , should be 2 E.U. for a substance already possessed of rotational freedom, as found for t-butyl chloride.

Relationships among the entropies of fusion of structurally similar compounds have also been sought,^{57, 58, 59} but without definite reference to structure. Melting points are often regarded as constitutive properties of organic compounds⁶⁰ and an empirical equation connecting them with molecular weights has been suggested.⁶¹ The linear relation between number of C atoms and melting points for odd or even numbers, respectively, of an homologous series, has been

⁵² Stuart, H. A. *Moleküllstruktur*. J. Springer. Berlin. 48. 1934.

⁵³ Herzfeld, K. F., & Mayer, M. G. *Phys. Rev.* 46: 995. 1934.

⁵⁴ Frenkel, Y., Todes, O., & Ismailow, S. *Acta Phys.-Chim. U. S. S. R.* 1: 97. 1934. Frenkel, Y. *Acta Phys.-Chim. U. S. S. R.* 4: 341. 1936.

⁵⁵ Hirschfelder, J., Stevenson, D., & Eyring, H. *Jour. Chem. Phys.* 5: 896. 1937.

⁵⁶ For other cases see Mott, N. F., & Gurney, R. W. *Trans. Faraday Soc.* 35: 364. 1939. Lennard-Jones, J. E., & Devonshire, A. F. *Proc. Roy. Soc. London. A* 170: 464. 1939.

⁵⁷ Walden, P. *Zeit. Elektrochem.* 14: 713. 1908.

⁵⁸ Kordes, E. *Zeit. Anorg. Allgem. Chem.* 160: 67. 1927.

⁵⁹ Clusius, K. *Zeit. Elektrochem.* 39: 598. 1933.

⁶⁰ Gilman, H. *Organic Chemistry, An Advanced Treatise*. John Wiley & Sons, Inc. Chap. 20. New York. 1938.

⁶¹ Austin, J. B. *Jour. Am. Chem. Soc.* 52: 1049. 1930.

long known,⁹⁰ and it has been observed without explanation that isomers often deviate from such straight lines.

Thermal and viscosity data are listed in TABLE 4, under the headings M.P., melting point; ΔS_f , entropy of fusion; T.P., transition point; ΔS_t , entropy of transition; $\Delta t^{\circ}I_{eq}$, liquid interval under normal conditions, in degrees; ΔE_{vis} , the activation energy for viscous flow, in calories per mole; and ΔE_{vap} , the energy of vaporization in calories per mole.

TABLE 4

Substance	M.P., °K.	ΔS_f	T.P., °K.	ΔS_t	$\Delta t^{\circ}I_{eq.}$	ΔE_{vis} cal/mole	$\frac{\Delta E_{vap}}{\Delta E_{vis}}$
CH ₃ Br	180.1				96.2		
Cl ₂ H ₅ Br	155.1	9.03			156.0		
n-C ₆ H ₁₃ Br	163.1	9.56			181.0	1770	
i-C ₄ H ₉ Br	184.1				148.4	1770	3.5
n-C ₄ H ₉ Br	160.4	13.5			214.3		
i-C ₄ H ₉ Br	155.1	(3.87) ^a			209.6	1990	3.5
t-C ₄ H ₉ Br	255.4	low	{ 231.7 ^b low 208.9 high }		90.9	2680	2.4
n-C ₆ H ₁₁ Br	185.1	18.6			216.7		
i-C ₆ H ₁₁ Br	161.1				232.6	1760	4.3
n-C ₆ H ₁₃ Br	188.1	23.0			241.0		
OCl ₄	250.3	2.30	255.5	4.79	99.9	2500	2.7
t-C ₄ H ₉ Cl	248.5	2.0	219.3 ^b		75.2	2490	2.4
OBr ₄	366.7	0.79	319.4	1.68	95.9		
C ₂ Cl ₆ ^c	459.9	4.3	344.2	7.3	0		
O(CH ₃) ₄	253.1				29.5		
C ₂ (CH ₃) ₆	377.1	4.51	148.1	3.24	2.8		
(CH ₃) ₂ C(OH) ₂ :CH ₂ ^d	158.4	1.65	124.9	8.32	155.9		
(OH) ₂ CH(OH) ₂ :CH ₂	154.0	13.8			200.1		
CH ₂ CH ₂ CH(C ₂ H ₅) ₂ CH ₂ CH ₂	154.3	14.7			212.1		
CH ₂ O(CH ₃) ₂ (CH ₂) ₂ CH ₂	148.1	9.5			204.5		
(CH ₃) ₂ CHCH ₂ CH(CH ₃) ₂	152.5	10.5			204.2		
CH ₂ OH ₂ C(CH ₃) ₂ CH ₂ CH ₂	138.2	12.2			221.0		
(CH ₃) ₂ C(OCH ₂ CH ₃) ₂	247.7	2.14	121.0	4.60	106.3		

^a Value probably low because of glass formation. Timmermans, J. Bull. Soc. Chim. Belg. 44: 17. 1935.

^b Reference 65.

^c Parijs, S. Zeit. Anorg. Allgem. Chem. 226: 424. 1936.

^d Kennedy, W. D., Shomate, C. H., & Parks, G. S. Jour. Am. Chem. Soc. 60: 1507. 1938. b. p.'s from Edgar, G., Calingaert, G., & Marker, R. E. Jour. Am. Chem. Soc. 51: 1483. 1929. Huffman, H. M., Parks, G. S., & Thomas, S. B. Jour. Am. Chem. Soc. 52: 3241. 1930.

FIGURE 13 demonstrates, as shown by the data of TABLE 4, that nearly spherical molecules, which have rotational transitions in the solid, melt far above the "normal" temperature, as t-butyl bromide melts at 255.4° K, 95° above n-butyl bromide. Further, those molecules of somewhat lower symmetry, which can still acquire some rotational freedom before melting, like that found above for i-propyl bromide, will still melt significantly higher and with correspondingly

lower entropies of fusion than their non-rotating isomers. This agrees with the position of i-propyl bromide on FIGURE 13, and, conversely, the normal location of the n-amyl bromide point confirms the dielectric results in deciding against rotational freedom in the solid.

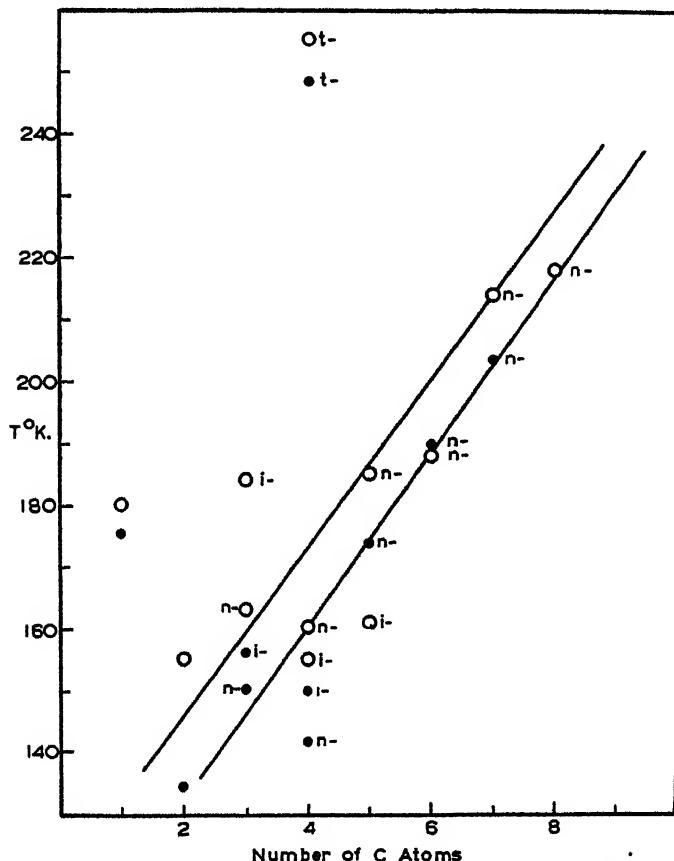


FIGURE 18. Graph of Melting Point against Number of Carbon Atoms in Alkyl Chlorides and Bromides. (Filled circles represent chlorides, hollow circles bromides.)

The abnormal positions of methyl chloride and bromide probably do not mean exceptional librational freedom, although rotation about the $C-X$ axis would not be detected by dielectric measurements, which, for methyl bromide, yield a solid state polarization only slightly higher than the molecular refraction.⁹² However, the

⁹² Morgan, S. O., & Lowry, H. H. Jour. Phys. Chem. 34: 2385. 1930.

methyl halide lattices contain a high concentration of dipoles, and might be expected to have greater lattice energies than the higher homologues, in which the dipoles are diluted by the hydrocarbon residues.

The following proposals indicate agreement between the thermodynamic and mechanical interpretations of melting properties. The trend toward equilibrium in a system under given conditions involves an increase in entropy while maintaining the minimum free energy feasible. At a given temperature, a solid can acquire entropy with small change of free energy by going over into the liquid, so melting occurs, and, at a still higher temperature, the boiling point, complete transition to the gas, the most disordered and hence highest entropy state of all, takes place. If, however, the symmetry of a molecule is such that its interaction with neighbors in a lattice does not vary much with its orientation, thermal agitation may give it sufficient kinetic energy to acquire, by means of a polymorphic transition, or simply a gradual loosening, a degree of rotational freedom approaching that in the liquid. A satisfying degree of disorder, or increase in entropy has been introduced, and the actual melting can be economically postponed to a higher temperature, when the tendency to increase the entropy still further, by gaining translational freedom, finally causes liquefaction. That a lattice whose molecules possess orientational freedom should have a reduced entropy of fusion is shown in another way by:

$$\Delta S_f = R \ln \frac{W_1}{W_s}, \text{ where}$$

W_1 = number of ways of realizing the liquid state,

W_s = number of ways of realizing the solid state,

in which $R \ln 2 = \sim 1.4 \text{ cal/mol}^{\circ}$,³⁸ for example.

From their approximate obedience to Trouton's Rule, most normal liquids are found to have the same entropy of vaporization at their boiling points. Hence, from the above argument, one would expect those molecules which gain extra entropy in the solid, and thus melt higher than is normal, to have liquid ranges much shorter than similar but less symmetrical molecules behaving normally in the solid. This is borne out in TABLE 4, where n-butyl bromide has a liquid range of 214° and t-butyl bromide is a liquid for only 91° . Striking examples of very symmetrical molecules with solid transitions that gain so little entropy on melting that they go over into the vapor

³⁸ Oldham, J. W. H., & Ubbelohde, A. R. Trans. Faraday Soc. 35: 328. 1939.

almost immediately are hexamethylethane, with 2.8° of liquid interval, and hexachloroethane, which, like iodine, forms no liquid under normal conditions. X-ray studies have shown the onset of rotational freedom in these two molecules,^{94, 95} as their non-polarity precludes dielectric investigation. The recorded ΔS_f of hexachloroethane, which would be the entropy change on sublimation less that of vaporization is little more than half the entropy of transition, so a vanishingly small liquid range is suggested.

i-Butyl and i-amyl bromides, which vitrified readily, appear on FIGURE 13 to melt too low, possibly because of the efficiency of their packing in the liquid as indicated by glass formation. However, since the dielectric studies showed that the molecules could not rotate in the crystalline solids, it would appear that the conditions of a small change in internal energy accompanied by a large, desirable entropy increase would lower the melting temperature.

The t-butylethylene data are included to provide another example besides carbon tetrabromide of an entropy of fusion less than the 2 units to be expected from communal volume sharing. Although its liquid range 160° is greater than those of the alkyl halides which have transitions, it is still a bit shorter than that of the unbranched butylethylene. The highly symmetrical hydrocarbon adamantine may also be expected to rotate in the solid, and possesses an elevated melting point in conformity with the theory.

Data for six of the isomeric heptanes, at the bottom of TABLE 4, further illustrate the generalizations. As the symmetry of the molecule finally reaches the maximum possible in 2, 2, 3 trimethylbutane, the melting point rises from a normal one around 150° K, to 247.7° K, and correspondingly, the liquid interval drops from an average of over 200° to 106.3° . Parks and Huffman⁹⁶ observed that increased branching in a paraffin hydrocarbon leads to a marked decrease in ΔS_f . This behavior is understandable in the light of the present results, where the additional symmetry to be derived from branching may produce librational freedom in the solid.

The applicability of the explanation of the melting point variation among isomers may be further examined in the camphor compounds. Camphor has a rotational transition in the solid state,^{97, 98} and is a symmetrical molecule, as may be seen from kinetic theory models.

⁹⁴ West, C. D. *Zeit. Krist.* **A 88**: 195. 1934.

⁹⁵ Finbak, C. *Chemical Abstracts*. **32**: 1996. 1938. *Tids. Kjemi Bergvesen* **17**: 9. 1937.

⁹⁶ Parks, G. S., & Huffman, H. M. *Ind. Eng. Chem.* **23**: 1138. 1931.

⁹⁷ White, A. H., & Morgan, S. O. *Jour. Am. Chem. Soc.* **57**: 2078. 1935.

If it is made unsymmetrical and the solid transition thus eliminated by moving the two methyl groups from the inner bridge to the outer ring, fenchone results, which should, by the above concept, have a lower melting point and a longer liquid interval than camphor. This is well confirmed by the comparison: d-camphor, m.p. 451.8° K, liquid interval, 30.4°; fenchone, m.p. 278.1° K, liquid interval, 188°. Similar considerations may be extended through the camphor series, where other rotational transitions are known.⁶⁷ Hence, the size of the molecules involved would not seem to limit the validity of the principle. Many similar cases amplify the principle, but only a final example previously discussed as a nearly spherical molecule will be cited: the melting point of t-butyl iodide is postponed to 234.9° K by its transition (see FIGURE 12); it fuses 65.3° higher than n-butyl iodide, m.p. 169.6° K; 55.3° higher than i-butyl iodide, m.p. 179.6° K; and 65.8° higher than sec.-butyl iodide, m.p. 169.1° K.

TABLE 4 also contains some viscosity data to be interpreted as discussed in the section on relaxation mechanism. That is, molecules so symmetrical that they behave as spheres will show a $\Delta E_{vap}/\Delta E_{vis}$ ratio of 2.5-3. The ΔE_{vap} values are from $\Delta H_{vap} - (pV)_{vap}$. The ratios are very approximate, but agree with the stereochemistry postulated for the fusion principle.

It may thus be concluded from fusion data, coupled with dielectric and crystallographic results, that members of a homologous aliphatic series (including derivatives and isomers) melting above the normal temperatures indicated by the graph of melting point against number of carbon atoms, or comparatively high melting members in any series have gained entropy by rotational freedom in the solid. Hence, since fewer than the normal degrees of freedom are acquired on fusion, these compounds will have abnormally short liquid intervals. Viscosity results agree with the other structural data in assigning this behavior to symmetrical molecules.

The writers wish to express their gratitude to the Editor of the Journal of the American Chemical Society for permission to reprint drawings previously published in the Journal.

SUMMARY

Dielectric constant and dipole loss measurements over a range of temperature and frequency on liquid and solid i-propyl bromide reveal a small hindered rotational freedom of the molecules in the anisotropic molecular lattice over a 40° interval below the freezing

point. Similar studies on n-amyl bromide show that its monotropic transition, which results in a melting point 5.9° higher than the freezing point, does not involve detectable rotational freedom.

Dielectric and optical examinations of i-butyl and i-amyl bromides disclose vitrification on cooling, followed by controlled crystallization at low temperatures. Generalizations as to glass formation have followed from these and other data. Molecules which pack efficiently in the liquid but do not form isotropic lattices should tend to vitrify. The temperature of narrowest distribution of relaxation times may indicate the point at which the vitreous aggregates begin to dissociate. Detailed dispersion data on the glasses have been used in the application of both the Stokes' law and absolute reaction rate theories to show that the inner friction constant in the usual Debye relaxation formula must be considered in terms of a mechanism involving higher activation energies than those obtaining in ordinary viscous flow. The cooperative nature of orientation processes has been emphasized by activation entropy calculations.

Polymorphism has been discovered and studied by thermal, optical and dielectric methods in t-butyl chloride, bromide and iodide. Dispersion, hysteresis and supercooling effects have been interpreted in terms of possible molecular processes.

The melting points, heats and entropies of fusion, and liquid intervals of alkyl halides and similar related molecules have been correlated with the shapes and mobilities of their units in the crystals. What is believed to be a new general principle of the melting behavior of symmetrical molecules has been proposed.

